Models and Inference for Complex Data with Applications in Nuclear Non-Proliferation and Microbial Systems

by

Haonan Zhu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in The University of Michigan
2023

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Dr. Andre Goncalves, Lawrence Livermore National Laboratory
Professor Xiaoxia Lin
Assistant Professor Qing Qu

Haonan Zhu
haonan@umich.edu
ORCID iD: 0000-0002-4239-4824

In Memory of my grandfather Rugen Zhu

ACKNOWLEDGEMENTS

When I first came to Ann Arbor in 2012, I would have never imagined this ends up becoming my home for the next 11 years. I am forever grateful for all the wonderful friends and colleagues I have crossed-paths with, and your kindness really made my time in Ann Arbor a memorable one.

I want to first thank my advisor Prof. Alfred Hero for giving me an opportunity to embark a research career in statistics and machine learning. A lot of people talk about interdisciplinary research, but few like you lives and breaths a life embodying the very word. I am always amazed by your ability to connect with people across fields ranging from plasma physics, social science to pure mathematics. Your devotion to research and helping others will always be an inspiration for me as a researcher and foremost a human being. I would like to extend my gratitude to the rest of my committee members who all offer valuable feedback on this thesis. To Dr. Andre Goncalves, thank you for being an amazing mentor to me during my internship at Lawrence Livermore National Laboratory, and I have benefited so much from your guidance and creativity in research. To Prof. Nina Lin, thank you for serving on my committee and challenging me to think about ways my research can be beneficial to the biology community. While there is still more progress to be made, thank you for your willingness to engage in conversations to bring two different disciplines together. To Prof. Qing Qu,

thank you for letting me work as a GSI with you for two semesters, and I would not have gained interests in optimization on Riemannian manifold without the teaching experience.

There are a couple more people I would like to thank in helping me to grow as a researcher. I would like to thank Prof. Yoann Altmann and Prof. Angela Di Fulvio for your assistance in my first major research project in graduate school. Thanks for all the faculty members in the signal processing and machine learning track: Prof. Raj Rao Nadakuditi, Prof. Laura Balzano, Prof. Clayton Scott and Prof. Jeff Fessler. I have taken at least one course from each of you and benefited significantly from your field of expertise. I was fortunate enough to be one of the last students taught by prof. Demosthenis Teneketzis before your retirement, and I want to thank you for your kindness during your office hours to talk about both research and life. I would like to thanks all the teaching staffs and students from EECS.564, EECS.545, EECS.559 and EECS.553, I have learned more about the foundation of machine learning through the experience. I would like to thank all the labmates I have the privilege to work and share an office space with. In particular, I would like to thank Dr. Salimeh Yasaei Sekeh, Dr. Morteza Noshad, Dr. Elizabeth Hou, Dr. Mayank Baranwal, Dr. Benoit Dufumier, Dr. Wayne Wang, Dr. Neophytos Charalambides, Dr. Mehmet Aktukmak, Zeyu Sun and Robert Malinas for all the discussion we had about mathematics and statistics. I would like to thank the DSSI program of Lawrence Livermore National Laboratory to give me an opportunity to work with an amazing interdisciplinary team led by Dr. Nicholas Be. Thanks for the program coordinator Jen and Nisha to make this internship as enriching as possible. Thanks for the machine learning team

including Dr. Andre Goncalves, Dr. Hiranmayi Ranganathana and Dr. Camilo Valdes, and our weekly discussions have always been a enjoyable time to find news ways to work with the microbiome data.

During my time in Ann Arbor, there is no other place that has left more impact on me than Harvest Mission Community Church. As pastors often share church is not a building but where God's people are. I am foremost grateful for the church community that has helped me to grow in my relationship with God on a daily basis. I want to thank all my past life group leaders, in particular Varoot for your patience and wisdom to allow me to understand grace for the first time, Ben for your years of friendship and brotherhood to walk with me through all the wild tail end of teenager years and more, Dr. Hong Yoon Kim for being an inspiration for me to love people, God and maybe math. I am grateful for the two life stage ministries I was part of: the Global Access ministry and Impact ministry. For GA has always been a home away from home for me, and Impact has really been a place where I was challenged to seek the welfare of the city where I am at. In addition, I would like to thank servant team (former helps team) that really taught me how to love the church in all the tangible ways possible.

I am thankful for all the friendships that supported me over the years. I would like to thank Timothy Wong for your steady commitment as a friend, and thank you for being a friend that I can discuss life, theology and stupid jokes with. Thanks for Chengcheng Zhu and Brian Purnomo for allowing me to drive you around randomly and share lives together, and your assistance during the past two months definitely contributed to my efforts to finish this thesis. Thanks for my friend Stanley Wang to be a brother I can always count on in time of emergencies

while being a partner in crimes. Thanks for Hee Sung Kim for being so far the only close friend ends up returning to Ann Arbor, and I am looking forward to all the amazing things you, Dorcas and Yuna will accomplish together. I want to thank my fellow graduate school friend James Tan for the partnership we shared in past 5 years to serve and love those around us, and your heart for people has really been inspiring for me. There are so many more individuals that I wish I could elaborate on their impact and friendship in my life but cannot for sake of space: Kexin Li, Jiacheng Lu, Yu Feng, Guang Sun, Junjie Yu, Qingwei Zhang, Qiming Yu, Zhipeng Wang, Qingzhou He, Eric Kwong, Jianghao Lu, Dr. Xingjian Lai, Jessica Kim, Alex Kim, Shanice Lau, Bingqing Zu, Erica Leung, Ping Khoo Lee, Hangil Lee, Dr. Nancy Wu, Dr. Sam Chen, Dr. Anita Li, Dr. Peter Li, Dr. Joshua Kammeraad, Allen Li, David Lee, David Chang, Torre Puckett, Zach Fritts, Dr. Hannah Abraham, Beatrix Yan, Dr. Grace Haeun Lee, Dr. Elaine Liu, Daniel Park, Alex Shen, Grace Chen, Esther Lee, Qiran Li, Elizabeth Choe, Angelica Li, Kevin Chang, Qiwei Lin, Jason Liu and so many others.

Regarding my family, I want to say thank you for your unconditional support for me to chase all the wildest dreams. Thanks for your commitment to stay in touch while we are in three different time zones throughout my PhD program, and forgave me all the time I failed to wake up on time. I am thankful for my parents, Aiqin Fang and Yigong Zhu, your belief in education is what gets me started. Thanks my sister Chaonan Zhu for being a role model for me from start. I am looking forward the day when we will be united in one geographic location.

At last, I own all my gratitude and worship to my lord and savior Jesus Christ. Through all my highs and lows, you have remained the one constant that always

draws me back to you. All in all, your grace is sufficient in my weakness, and your steadfast love never ceases even when I was not looking for it.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

With recent advances in science and technology, researchers are often provided with unprecedented amounts of complex data to analyze. The structure of the data, due to being high dimensional, discrete, incomplete, extrapolating meaningful information from these data requires models that incorporate knowledge about the underlying systems and implement efficient computation methods. In this thesis, we have developed statistical models and inference algorithms (using Monte Carlo methods and optimization methods) comprehensively undertaken their performance analysis for several complex problem domains. These domains include: inverse problems in radiation detection; data fusion and classification in high dimensional microbiome studies; and contingency table analysis for inferring voting patterns in election polling data with missing information.

Specifically, Chapter II describes a hierarchical Bayesian model and state-of-art Monte Carlo sampling method to solve the unfolding problem, i.e., to estimate the spectrum of an unknown neutron source from the data detected by an organic scintillator. The proposed approach is compared to three existing methods using simulated data to enable controlled benchmarks. Our results show that the proposed method has competitive unfolding performance compared to existing approaches in terms of accuracy and robustness against limited detection events, while requiring less user supervision. The proposed method also provides additional posterior confidence measures.

Chapter III develops a Bayesian graphical model for fusing disparate types of count data. The motivating application is the study of bacterial communities from diverse high-dimensional features. We introduce a flexible multinomial-Gaussian generative model for jointly modeling such count data. We present a computationally scalable variational Expectation-Maximization (EM) algorithm for inferring the latent variables and the parameters of the model. The inferred latent variables provide a common dimensionality reduction for visualizing the data. In addition to simulation studies that demonstrate the variational EM procedure, we apply our model to a bacterial microbiome dataset.

Chapter IV proposes a hierarchical Bayesian multitask learning model that is applicable to the general multitask binary classification learning problem where the model assumes a shared sparsity structure across different tasks. We derive a computational efficient inference algorithm based on variational inference to approximate the posterior distribution. We demonstrate promises of the new approach on multiple synthetic datasets and a real world microbiome dataset in comparison with other benchmark methods.

Chapter V introduces an exact model with minimal assumptions for the transition matrix recovery problem, where we are given multiple two-way contingency tables with known margin sums but missing inner cells. We propose three valid approximations of the exact model and a novel Riemannian gradient algorithm to obtain the Maximum Likelihood Estimators (MLE) of the transition matrix. The proposed methods are applied to a synthetic dataset and a real world dataset from the New Zealand general election. Our simulation studies show the scope when those approximations apply. A further clustering analysis using the esti-

mated stochastic matrices across different electorate districts is able to identify communities that are reflective of the demographics of New Zealand.

CHAPTER I

Introduction

With recent advances in science and technology, researchers are often provided with unprecedented amounts of complex data to analyze. The structure of the data, due to being high dimensional, discrete, incomplete, extrapolating meaningful information from these data requires models that incorporate knowledge about the underlying systems and implement efficient computation methods. In this thesis, we have developed statistical models and inference algorithms (using Monte Carlo methods and optimization methods) comprehensively undertaken their performance analysis for several complex problem domains. These domains include: inverse problems in radiation detection; data fusion and classification in high dimensional microbiome studies; and contingency table analysis for inferring voting patterns in election polling data with missing information.

1.1 Statistical Modeling for Complex Systems

A general systems perspective provides context for the research presented in this thesis. In the study of traditional physical systems, once we have a complete characterization of the governing equations of the system, we can predict the outputs of the system given the initial states or inputs of the system. This task is

called a forward problem, and can be visualized abstractly in Fig. 1.1:

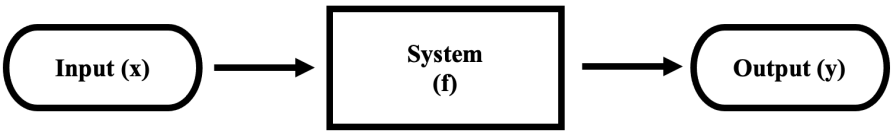


Figure 1.1: General Input-Output System

In contrast, the process of using knowledge about the system to determine the input associated with a given output is called an inverse problem [1]. This problem arises in applications where the experimenter only has access to measurements of the output and wishes to infer the input, or some specified properties of the input. For example, in nuclear radiation detection and identification that is treated in Chapter II, it may be of interest to discriminate between possible isotopes that generated the neutrons, the neutron source. The objective could be to classify the source as benign, e.g. trace medical isotopes such as Technitium 99m or Iodine 131, or suspiciious, e.g. a special nuclear material (SNM) containing fissionable isotopes like uranium-233, uranium-235, or plutonium-239, introduced with malicious intent. The neutron energy spectrum provides the key discriminant for deciding between benign vs malicious sources, and the detector has a system transfer function that translates the unobserved spectrum of an incident neutron at its input to the observed burst of light at the output. The neutron spectrum can only be extracted through deconvolution of the measured light output spectrum and the response functions of the detector to monoenergetic neutrons. Due to attenuation, scattering, and background noise, it is necessary to use all available information about the system in order to accurately recover the neutron spectrum.

To counter this loss of information, in Chapter II we adopt a Bayesian perspective of this inverse problem that introduces Bayesian priors in order to incorporate prior knowledge about the inputs into the inference [2].

Data fusion, where the same inputs are measured by different systems (e.g sensors), is the integrative process of aggregating and synthesizing information about the input across multiple sources. This process is applicable to problems where data from multiple sensors are readily available and each sensor only measures partial information about the inputs. For example, when studying microbial systems, one may aim to quantify how changes in environmental conditions affect microbial community profiles, with the goal of developing sensors based on these biotic indicators. A common way to obtain a global profile of a microbial community is to perform gene sequencing on a biological sample. In particular, RNA-Seq measures gene expression in a community by quantifying the number of times each gene transcript occurs in the pool of sequenced RNAs. Each microbial species in the community is represented by its own unique set of transcripts, i.e., its transcriptome, and fusing information from different transcriptomes yields the global profile of gene expression across all species in the community. This type of analysis is known as metatranscriptomics. Chapter III introduces a Bayesian graphical model for the metatranscriptomics problem to capture patterns of similarity between histograms of different species' gene expression without inter-species genome-to-genome mappings nor knowledge of inter-species transcriptomic pathway correspondences in reflection of condition changes.

In Multitask learning, observations of multiple inputs and outputs of related

systems (tasks) are observed, and the goal is to infer the unknown systems for future predictive tasks. This framework arises in applications where we have limited number of input-output pairs for each system. For example, in the study of human gut microbiomes, it may be of interest to perform health prediction based on human gut data. However, microbiome data presents two major challenges. First typical microbiome data lies in high feature dimension, i.e the number of microbes is significantly more than the number of samples available. Second, in health applications it is essential that machine learning models be interpretable and quantify uncertainty in their predictions. To address these challenges, in Chapter IV we deploy a Bayesian multitask learning framework with variational approximations to perform predictions jointly on multiple datasets, applying to gut microbiome prediction of health outcomes.

In Chapter V, we formulate the transition matrix recovery problem as estimation of a stochastic matrix of conditional probabilities from multiple experiments, where we are given multiple two-way contingency tables with known margin sums but missing inner cells. In this problem the margin sums of two categories of the contingency tables are the input-output pairs of the system governed by the transition matrix, and we are interested in recover the conditional probabilities that describes the system. This problem arises when data are only available at the population level instead of individual level due to either limitations of the measurements or privacy constraints. For example, in political science, exit polls often collect information on how voters voted in an election. They may also ask the voter how she voted in the last election or what party affiliations were held by the candidates she voted for in a multi-category election, e.g., an election for

state legislature, federal congress, and presidential candidates. Exit poll voting data is often separately aggregated according to electoral districts and individual level voter-specific cross-tabulated data may not be reported. In this case only marginal data per district is available and the results in Chapter V can be applied to recover the conditional probability (transition) matrices associated with voter choices across the election categories.

1.2 Outline and Contributions

This section lists the chapters and corresponding contributions in this thesis. Each chapter aims to be a self contained exposition on a specific topic.

Chapter II describes a hierarchical Bayesian model and state-of-art Monte Carlo sampling method to solve the unfolding problem, i.e., to estimate the spectrum of an unknown neutron source from the data detected by an organic scintillator. The proposed approach is compared to three existing methods using simulated data to enable controlled benchmarks. We consider three sets of detector responses. One set corresponds to a 2.5 MeV monoenergetic neutron source and two sets are associated with (energy-wise) continuous neutron sources (^{252}Cf and ^{241}AmBe). Our results show that the proposed method has similar or better unfolding performance compared to other iterative or Tikhonov regularization-based approaches in terms of accuracy and robustness against limited detection events, while requiring less user supervision. The proposed method also provides a posteriori confidence measures, which offers additional information regarding the uncertainty of the measurements and the extracted information. This chapter is based on the work of [3] that was published in *IEEE Transactions on Nuclear Science*.

Chapter III [1] develops a Bayesian graphical model for fusing disparate types of count data. The motivating application is the study of bacterial communities from diverse high-dimensional features. We introduce a flexible multinomial-Gaussian generative model for jointly modeling such count data. This latent variable model jointly characterizes the observed data through a common multivariate Gaussian latent space that parameterizes the set of multinomial probabilities of the transcriptome counts. The covariance matrix of the latent variables induces a covariance matrix of co-dependencies between all the transcripts, effectively fusing multiple data sources. We present a computationally scalable variational Expectation-Maximization (EM) algorithm for inferring the latent variables and the parameters of the model. The inferred latent variables provide a common dimensionality reduction for visualizing the data and the inferred parameters provide a predictive posterior distribution. In addition to simulation studies that demonstrate the variational EM procedure, we apply our model to a bacterial microbiome dataset. This chapter is based on the work of [4] that is going to be published in *IEEE Transactions on Signal Processing*.

Chapter IV [2] proposes a hierarchical Bayesian multitask learning model that is applicable to the general multitask binary classification learning problem where the model assumes a shared sparsity structure across different tasks. We derive a computational efficient inference algorithm based on variational inference to approximate the posterior distribution. We demonstrate promises of the new approach on multiple synthetic datasets and a real world microbiome dataset pooled from multiple distinct studies in comparison with other benchmark methods.

[1]This work was partially supported by grants from ARO W911NF-19-102 and DOE DE-NA0003921.

[2]This work was partially supported by the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 with IM release number LLNL-MI-853606. It was also partially supported by the US Army Research Office under grant number W911NF1910269.

Our results in synthetic datasets show that the proposed approach has superior support recovery property when the underlying regression coefficients share a common sparsity structure across different tasks. Though our experiments on real world dataset do not show improvement of the proposed model in terms of prediction metrics due to the pooled datasets are heterogeneous (i.e different experimental objectives, laboratory setups, sequencing equipments, patient demographics etc.), we demonstrate the utility of the method to extract informative taxons while providing well-calibrated predictions with uncertainty quantification.

Chapter V introduces an exact model with minimal assumptions for the transition matrix recovery problem, where we are given multiple two-way contingency tables with known margin sums but missing inner cells. We propose three valid approximations of the exact model and a novel Riemannian gradient algorithm with Polyak adaptive step size to obtain the Maximum Likelihood Estimators (MLE) of the transition matrix. The proposed methods are applied to a synthetic dataset and a real world dataset from the New Zealand general election. Our simulation studies show the scope when those approximations apply. A further clustering analysis using the estimated stochastic matrices across different electorate districts is able to identify communities that are reflective of the demographics of New Zealand.

CHAPTER II

A Hierarchical Bayesian Approach to Neutron Spectrum Unfolding with Organic Scintillators

2.1 Introduction

Two main reactions are exploited in neutron detection: scattering on a light nucleus or capture on elements such as ^{6}Li, ^{10}B or ^{3}He. Thermal neutrons (0.025 eV) are preferentially detected via capture reactions because the aforementioned elements exhibit high cross-sections for thermal neutron absorption. Conversely, fast neutrons are detected via scattering reactions on light elements, such as hydrogen and deuterium. The detection of fast neutrons, such as those emitted by SNMs, involves directly exploiting inelastic and elastic scattering reactions, without the need to moderate the source neutrons. Organic scintillators are typically hydrocarbon compounds and detect neutrons via elastic and inelastic scattering reactions on hydrogen nuclei. The energy deposited by scattered proton recoils depends on the scattering angle and it ranges from zero up to the neutron maximum energy. The intensity of light pulses produced by the scintillator is correlated to the energy deposited by the recoil protons [5]. This light production mechanism allows partial retention of the energy of the impinging neutrons, however, the correlation between the energy of the impinging neutron and the

light pulse produced is weak, and therefore deriving the neutron spectrum from the measured data is particularly challenging.

Finding the energy spectrum of the neutrons impinging on an organic scintillator from its light output response is an ill-posed problem, which often admits multiple solutions [6]. This problem is traditionally addressed using so-called unfolding algorithms, which aim at recovering the spectrum that is most likely to have produced the given measured response. Accurate unfolding and spectrometry are critical in several applications, such as radiation protection [7], nuclear physics [8], nonproliferation [9] and safeguards [10]. In safeguards, nonproliferation, and decommissioning applications, accurately discriminating between different neutron sources, such as those based on (alpha, n) reactions and those based on fission, would be a valuable tool when characterizing neutron-emitting samples of unknown composition.

2.2 Background

2.2.1 Organic Scintillator Response and Monte Carlo Simulation

Scintillators emit light upon interaction with ionizing radiation. Organic scintillators are compounds of hydrogen and carbon, and are suitable to detect fast neutrons. Neutron elastic scatter on a hydrogen nucleus produces a scattered neutron and a recoil proton. In the energy range of interest ($<$ 20 MeV neutrons), it can be assumed that the recoil proton deposits all its energy within a detector of practical size, e.g. 7.62-cm diam. by 7.62-cm length. The light output response is approximately linear with the energy deposited by electrons, E_e, with energy above approximately 40 keV [11]. Therefore, the detector light output is conveniently expressed in terms of electron light output (*ee*: electron-equivalent units). In

practice, the upper edge of the known Compton electron distribution produced by a monoenergetic gamma-ray source, e.g. ^{137}Cs, provides a suitable calibration point, commonly referred to as the Compton edge, V_{CE}. The light output in electron equivalent units (y_{ee}) is therefore calculated at any pulse height voltage V as in Eqn. (2.1).

$$y_{ee} = \frac{\bar{E}_{ee}}{V_{CE}} V. \tag{2.1}$$

In equation (2.1), $\bar{E}_{ee}$ is the maximum energy deposited by a Compton-recoil electron, in electron-equivalent energy units. Conversely, the light output response to charged particles heavier than electrons, like neutron-produced recoil protons, is not linear with the energy deposited. Throughout this paper, y identifies the light output in electron-equivalent energy units, e.g., keV_{ee}. A widely accepted set of models which semi-empirically describes the dependence of the light output y with the proton energy deposited E_p and the energy deposited-per-unit-length dE_p/dx was first introduced by Birks [5] and is reported in Eqn. (2.2) below

$$y(E'_p) = \int_0^{E'_p} \frac{S\, dE_p}{(1 + k_B\, dE_p/dx)}. \tag{2.2}$$

Equation (2.2) is the integral over energy of Eqn. (3) in the paper by Brooks *et al.* [12]. In Eqn. (2.2), S is the scintillation efficiency, in MeV_{ee}, and k_B is a material-dependent constant, in $g/MeVcm^2$, often referred to as the Birks' coefficient [12]. We simulated the pulse height distributions, i.e. light output spectra, of a 7.62-cm diam by 7.62 length EJ-309 detector in response to monoenergetic neutrons, for 500 evenly distributed neutron sources with energy between 0.1 MeV to 20 MeV, using MCNPX-PoliMi [13]. We used MPPost, a MCNPX-PoliMi post-processing code, to obtain the light output spectrum , i.e. the frequency of occurrence of

pulse amplitudes in a given measurement time [14]. An enhanced version of MPPost allows the use of the semi-empirical model in equation (2.2) to generate the detector-specific light output spectrum [15]. For EJ-309, the coefficients S and k_B that we used are 2.277 MeVee/MeV and 33.84 g/MeV cm^2, respectively [15]. The software also applies a Gaussian smear to account for the detector's energy resolution. The energy resolution function that we implemented was measured by Enqvist et al. [16] for the type of detector under investigation and is reported in Eqn. (2.3), where $a = 0.113 \pm 0.007$, $b = 0.065 \pm 0.011 MeV^2$, and $c = 0.060 \pm 0.005 MeV$.

$$(\Delta E/E) = (\sqrt{a^2 + b^2/E + (c/E)^2}) \tag{2.3}$$

Fig. 2.1 shows the simulated light output spectra produced by irradiation with selected mono-energetic neutron sources between 0.5 MeV and 5 MeV.

The energy deposited in the detector by recoil protons E_p after elastic collision with neutrons of energy E depends on the scattering angle of the charged recoil in the laboratory system of reference: θ (see Eqn. (2.4)).

$$E_p = \frac{4A}{(1+A)^2}\, cos^2\theta\, E \tag{2.4}$$

In the elastic scattering kinematics equation (Eqn. (2.4)), A is the mass number of the target nucleus (A=1 for ^{1}H). Monoenergetic neutrons can thus produce proton recoils in the energy range from $E_{pmax} = E$, when $\theta = 0$, to zero, when $\theta = \frac{\pi}{2}$ and consequently light pulses with amplitude ranging form $y(E_{pmax})$ to 0. Note that in Fig. 2.1, the light output corresponding to the maximum energy deposited by proton recoils is identified by solid diamonds. We determined this

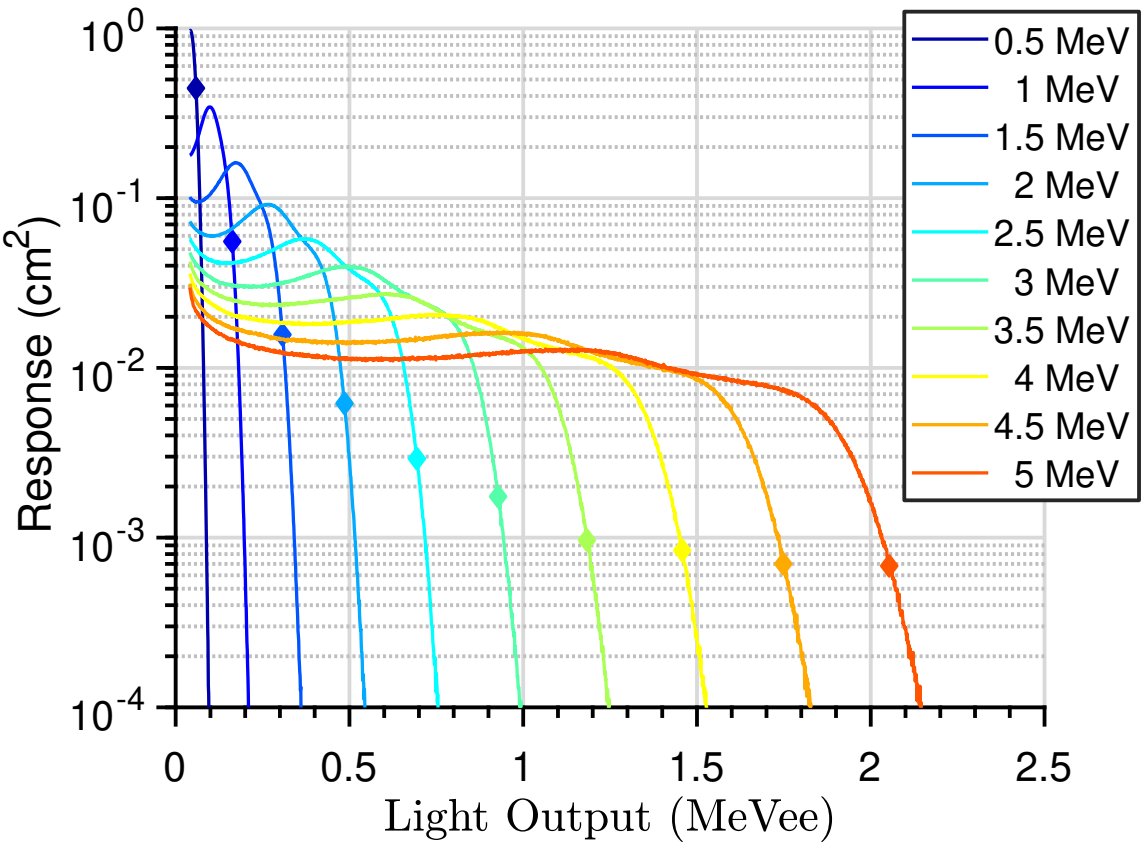


Figure 2.1: Simulated Response Functions for a 7.26 cm diam. by 7.26 cm length EJ-309 detector in response to monoenergetic neutrons in the 0.5-5 MeV range. Tshe solid diamonds show the light output corresponding to the maximum energy deposited.

light-output value as the minimum of the derivative of the upper edge of the light output spectrum, following the same method proposed by Kornilov and colleagues [17].

As in any spectroscopy-capable sensor, the number of counts at a given bin of the light output spectrum $y(E')$ (E' in *ee*) is given by the convolution of the detector response at that light output bin with the impinging neutron spectrum, as formalized in the next section (Eqn. (2.5)). Fig. 2.2 shows the process of spectrum unfolding for two monoenergetic neutron spectra on discretized data sets. One may notice that an ideal monoenergetic neutron spectrum is a linear transformation of one element of the canonical basis for the response matrix and

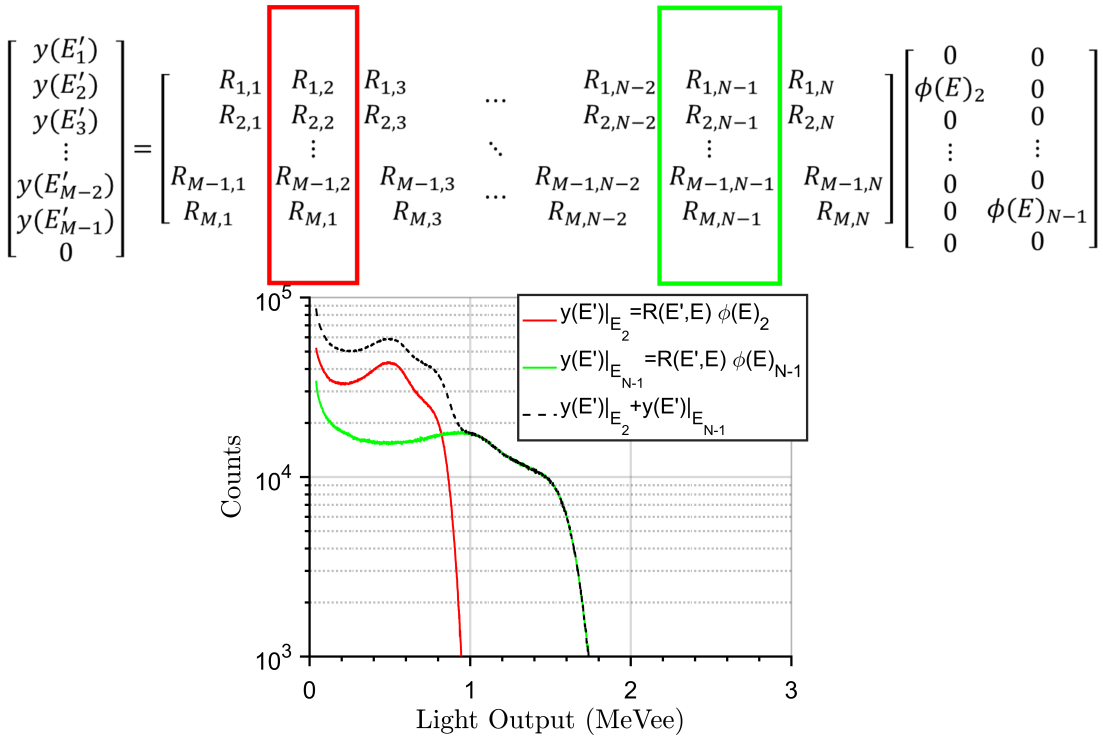


Figure 2.2: Example of the convolution between an ideal neutron spectrum with two energy peaks and the detector response matrix.

therefore selects only one corresponding light output response, i.e. column of the response matrix. For organic scintillation detectors, the number of neutron energy bins (N) is of the same order of magnitude as the number of light-output channels measured (M). In neutron spectroscopy, this case is usually referred to as multi-channel unfolding, as opposed to few-channel unfolding, where $M << N$. Few-channel unfolding applies to other types of detectors, e.g. Bonner spheres [18] and superheated emulsions [19]. The size of the response matrix used in this chapter is 600×149 (i.e., $M = 600$ and $N = 149$). These channel numbers correspond to a light output bin width of 0.001 MeVee, in the 0.01-6 MeVee light-output range, and a neutron energy bin width of 100 keV, in the 0.1-15 MeV energy range.

2.2.2 Discretized observation model

The detector response function is denoted by $R(E', E)$. More precisely, $R(E', E_0)$ is the light output spectrum (with E' in eVee) in response to a monoenergetic neutron of energy E_0. The light output and unknown neutron energy spectral fluence, i.e. the number of neutrons per unit area [20], also referred to as neutron spectra throughout this paper, are related through the following Fredholm integral equation [21, 22, 23, 24, 25]

$$y(E') = \int_0^\infty R(E', E)\phi(E)dE. \tag{2.5}$$

For numerical computation, Eqn. (2.5) can be approximated by the following linear equation

$$\boldsymbol{y} \approx \boldsymbol{R\phi}, \tag{2.6}$$

where $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_N]^T \in \mathbb{R}_+^N$ denotes the neutron spectrum discretized over N energy bins, $\boldsymbol{y} = [y_1, \ldots, y_M]^T \in \mathbb{R}_+^M$ is light output spectrum discretized over M bins and $\boldsymbol{R}$ is the $M \times N$ response matrix of the detector. Unfolding methods aim at recovering $\boldsymbol{\phi}$ from $\boldsymbol{y}$ such that Eqn. (2.6) is satisfied. However, they can differ by the similarity measures or likelihood functions used to compare $\boldsymbol{y}$ and $\boldsymbol{R\phi}$. A classical approach to matching $\boldsymbol{y}$ and $\boldsymbol{R\phi}$ consists of considering a quadratic similarity measure

$$||\boldsymbol{y} - \boldsymbol{R\phi}||_{\boldsymbol{\Sigma}}^2 = (\boldsymbol{y} - \boldsymbol{R\phi})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{R\phi}), \tag{2.7}$$

where the $M \times M$ matrix $\boldsymbol{\Sigma}$ relates to the characteristic of the measurement noise. If $\boldsymbol{\Sigma}$ is set to the identity matrix, Eqn. (2.7) reduces to the classical least-squares criterion $||\boldsymbol{y} - \boldsymbol{R\phi}||_2^2$ where $||\cdot||_2$ denotes the standard ℓ_2 norm. Recovering $\boldsymbol{\phi}$

using the criterion in Eqn. (2.7) implicitly assumes that $\boldsymbol{y}$ is a noisy version of $\boldsymbol{R\phi}$ corrupted by Gaussian noise with covariance matrix (proportional to) $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{y}|\boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{R\phi}, \boldsymbol{\Sigma}\right), \tag{2.8}$$

where $\boldsymbol{y}|\boldsymbol{\phi}$ reads "$\boldsymbol{y}$ given $\boldsymbol{\phi}$", $\sim$ reads "is distributed according to" and $\mathcal{N}\left(\boldsymbol{m}, \boldsymbol{\Sigma}\right)$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Indeed, it can be easily shown that minimizing (2.7) with respect to (w.r.t.) $\boldsymbol{\phi}$ is equivalent to maximizing the likelihood (2.8) w.r.t. $\boldsymbol{\phi}$, as will be discussed in the next section.

Since the acquisition process consists of detecting individual neutrons (discrete number of events within a given time period), it is reasonable to consider Poisson noise models. These models enable the consideration of the correlation between the mean (expected) detection rates and the variance of the observation noise. Moreover, such models are more suited for low counts (e.g. less than 10 per bin), as investigated in Section 2.4 where we consider scenarios with as few as 1 count per light output bin on average. The classical Poisson noise model assumes that the light output in the M energy bins are mutually independent and Poisson distributed. The resulting observation model becomes [26]

$$\boldsymbol{y}|\boldsymbol{\phi} \sim \mathcal{P}\left(\boldsymbol{R\phi}\right), \tag{2.9}$$

where $\mathcal{P}\left(\cdot\right)$ denotes the element-wise Poisson distribution, i.e., $\forall m, y_m|\boldsymbol{\phi} \sim \mathcal{P}\left(\boldsymbol{r}_{m,:}\boldsymbol{\phi}\right)$ with $\boldsymbol{r}_{m,:}$ the mth row of $\boldsymbol{R}$. Consequently, the likelihood of the observed light output spectrum $\boldsymbol{y}$ given the underlying neutron spectrum $\boldsymbol{\phi}$, denoted $f(\boldsymbol{y}|\boldsymbol{\phi})$ can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{\phi}) = \prod_{m=1}^{M} \frac{(\boldsymbol{r}_{m,:}\boldsymbol{\phi})^{y_m}}{y_m!} \exp\left[-\boldsymbol{r}_{m,:}\boldsymbol{\phi}\right]. \tag{2.10}$$

In this subsection, we have discussed how the unfolding problem can be formulated as a linear inverse problem and discussed two main noise observation models. In the next subsection, we review the primary existing unfolding methods and their relation with the observation models discussed above. These methods will then be used in Section 2.4 to assess the performance of the proposed approach.

2.2.3 Existing unfolding approaches

The first statistical approach to unfolding is a classical method for inverse problems and is referred to as Maximum Likelihood Estimation (MLE). MLE-based unfolding recovers the neutron spectrum by finding $\boldsymbol{\phi}$ that maximizes the likelihood function [27]. Maximizing the likelihood $f(\boldsymbol{y}|\boldsymbol{\phi})$ is equivalent to minimizing the negative log-likelihood, (which is often preferred for algorithmic stability since $-\log\left(f(\boldsymbol{y}|\boldsymbol{\phi})\right)$ is often a (nearly) quadratic function). Although we can consider as many MLE-based algorithms as likelihood models, we primarily focus on Gaussian and Poisson noise models here. More precisely, using an isotropic Gaussian noise model is equivalent to using a classical minimization of least square loss, while the Poisson model is preferred for counting data as discussed above. Under Poisson noise assumption, the log-likelihood reduces to

$$\begin{aligned}&\log(f(\boldsymbol{y}|\boldsymbol{\phi}))\\ &= \sum_{m=1}^{M} y_m \log(\boldsymbol{r}_{m,:}\boldsymbol{\phi}) - \log\left(y_m!\right) - \left(\boldsymbol{r}_{m,:}\boldsymbol{\phi}\right).\end{aligned} \tag{2.11}$$

Maximum likelihood estimation aims at recovering the unknown spectrum from the data only, i.e., without additional information), by inverting (or pseudo inverting) the response matrix and using a cost function accounting for the statistical properties on the observation noise. This is a simple inference strategy

but can provide poor results in the presence of noise, especially when the response matrix is ill-conditioned (as it is often the case in practice). Thus, maximum penalized likelihood estimation methods based on Poisson likelihood models have been proposed. Since we expect most of the unknown neutron spectra to be recovered are relatively smooth, it makes sense to add a regularization which reflects this prior belief. Here we chose a regularization term that promotes small second-order derivative (in the spectral dimension), which results in the following objective function to be minimized

$$\min_{\boldsymbol{\phi}\in\mathcal{R}_+^N} \sum_{m=1}^{M} -\log(f(\boldsymbol{y}|\boldsymbol{\phi})) + \lambda||\boldsymbol{L}\phi||_2^2, \tag{2.12}$$

where λ is a tuning parameter that controls the smoothness, $\log(f(\boldsymbol{y}|\boldsymbol{\phi}))$ is defined in (2.11) and $\boldsymbol{L}$ denote the discrete Laplace operator, which can be written as

$$\boldsymbol{L} = \begin{pmatrix} -2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & & & & \vdots \\ 0 & 1 & -2 & 1 & \ddots & & & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & -2 & 1 & 0 \\ \vdots & & & & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & -2 \end{pmatrix}. \tag{2.13}$$

There are multiple ways of solving the minimization problem in Eqn. (2.12), e.g., using Alternating Direction Method of Multipliers (ADMM) [28] as in Poisson image deconvolution by augmented Lagrangian (PIDAL) (see [29]) or using sequential Gaussian approximations of the Poisson likelihood [30]. Here, we chose the ADMM implementation presented in [29] for its simplicity and relatively low

computational cost. It is worth noting that the One-Step-Late (OSL) algorithm in [31, 32] is an alternative method to approximate the solution of Eqn. (2.12). Note that Eqn. (2.12) requires to select an appropriate value of λ, which will affect the quality of the solution. This point will be further discussed in Section 2.4.

Under the Gaussian noise model, the unfolded spectrum is a solution to the convex optimization problem as in (2.12) where $-\log(f(\boldsymbol{y}|\boldsymbol{\phi}))$ is replaced with the standard quadratic loss function $||\boldsymbol{y} - \boldsymbol{R}\boldsymbol{\phi}||_2^2$. The non-negativity constraints imposed on the unfolded spectrum prevent us from having a closed form solution, thus we applied an ADMM algorithm with L-curve method [33] to obtain the unfolded spectrum. This algorithm will be referred to Tik (Tikhonov Regularizer) in remainder of the paper.

Among the methods whose codes are available, we also used GRAVEL presented in [34, 35]. The iterative update rule of GRAVEL algorithm (at iteration $(k+1)$) is given by

$$\phi_n^{(k+1)} = \phi_n^{(k)} \exp\left(\frac{\sum_m W_{n,m}^{(k)} \log\left(\frac{y_m}{\boldsymbol{r}_{m,:}\boldsymbol{\phi}^{(k)}}\right)}{\sum_m W_{nm}^{(k)}}\right), \forall n, \tag{2.14}$$

where $\boldsymbol{\phi}^{(k)}$ is estimated neutron spectrum at iteration k, σ_m is an estimate of measurement error in the mth light output bin, $r_{m,n} = [\boldsymbol{R}]_{m,n}$ and

$$W_{n,m}^{(k)} = \frac{r_{m,n}\boldsymbol{\phi}_n^{(k)}}{\sum_i \left(r_{m,i}\boldsymbol{\phi}_i^{(k)}\right)} \frac{y_m^2}{\sigma_m^2} \tag{2.15}$$

GRAVEL allows the user to incorporate prior information, when available, as an a priori known default spectrum. We have used a flat spectrum for consistency with the other methods. Regardless of the type of source, a flat initial spectrum was used, whose boundaries are detailed in Table 2.1. The spectrum intensity had a negligible impact on the final results. The boundaries of the light output

spectra are reported in Table 2.1 and vary according to the simulated data. Light-output bins with a relative statistical error higher than 20% in the high-energy tail of the light output spectra were excluded. The uncertainty associated with the simulated bins was calculated as the square root of the counts. GRAVEL stopping criterion is either the user-defined chi-squared per degree of freedom (PDF) or the input maximum number of iterations (to stop the algorithm after a given number of iterations if the first criterion is not satisfied yet) [36]. In our case, the number of degrees of freedom is M and the chi-squared-PDF was set to one, while the maximum number of iterations was 6000. For the ^{252}Cf and ^{241}AmBe spectra (see Section 2.4), the algorithm reached the desired chi-squared PDF after few iterations (< 20), while the maximum number of iterations criterion was adopted for the monoenergetic spectrum, for which the relative fluctuation in the chi-squared PDF was below 0.0004%, after 6000 iterations. The GRAVEL parameters used in Section 2.4 are reported in Table 2.1.

Table 2.1 Specific parameters and settings used to unfold the neutron spectra in GRAVEL.

Parameters	^{241}AmBe	^{252}Cf	2.5 MeV
LO_{min}-LO_{max} (MeVee)	0.05-5.8	0.05-4.2	0.05-0.83
E_{min}-E_{max} (MeV)	0.5-15.0	0.5-15.0	0.5-3.0

MAXED is another unfolding computer program available within the UMG package [37]. MAXED applies the maximum entropy principle to the deconvolution of spectrometer data. The obtained results were similar to those calculated using GRAVEL, therefore MAXED was not included as an additional comparison methods.

2.3 Hierarchical Bayesian spectrum unfolding

2.3.1 Proposed Model

Bayesian methods have been previously proposed [21, 23, 26, 38, 39] in the context of spectrum unfolding. As mentioned earlier, they aim at regularizing ill-posed problems by incorporating a-priori information about $\boldsymbol{\phi}$ in a principled way. More precisely, such knowledge is incorporated through a so-called prior distribution $f(\boldsymbol{\phi}|\delta)$, parameterized by δ. The selection of the prior distribution $f(\boldsymbol{\phi}|\delta)$ is guided by the amount of prior information available and the induced algorithm complexity [26]. Moreover, the choice of this distribution can be crucial when the amount of information contained in the data in limited, e.g., in the presence of few observations and noisy data. While informative prior distributions will greatly improve the estimation performance if appropriately tailored, they will negatively impact the estimation performance if the data deviates from the the prior belief. In previous studies [21, 23], empirical Bayes methods were used, in which the prior distribution was built from previously acquired data. However, such methods perform poorly if the neutron spectrum to be recovered is not in agreement with the data-driven prior distribution. Bayes' theorem provides a formal way to combine our prior belief $f(\boldsymbol{\phi}|\delta)$ with the observations (through the likelihood $f(y|\boldsymbol{\phi})$) to obtain and exploit $f(\boldsymbol{\phi}|\boldsymbol{y}, \delta)$. This so-called posterior distribution is classically exploited using summary statistics, including various Bayesian point estimators such as the widely used maximum a posterior (MAP) estimator [21, 23] (which can also be seen as maximum penalized likelihood estimation) and posterior means (as in [26]) and a posteriori measures of uncertainty (e.g., confidence regions). However, the posterior distribution (e.g.

its mode or mean) can highly depend on the value of δ. A classical approach thus consists of incorporating this parameter in the estimation process by extending the Bayesian model and designing an additional prior distribution $f(\theta)$. Applying the Bayes' rule to that model leads to

$$f(\boldsymbol{\phi}, \delta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\phi})f(\boldsymbol{\phi}|\delta)f(\delta)}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\boldsymbol{\phi})f(\boldsymbol{\phi}|\delta)f(\delta), \tag{2.16}$$

where the posterior distribution $f(\boldsymbol{\phi}, \delta|y)$ summarizes the complete information available about (ϕ, δ), having observed $\boldsymbol{y}$.

In a similar fashion to the penalized likelihood method in (2.12), we choose to assume that the unknown neutron spectrum to be recovered presents smooth variations across neighboring energy bins. This is achieved by assigning $\boldsymbol{\phi}$ a truncated multivariate Gaussian distribution

$$\boldsymbol{\phi}|\delta \sim \mathcal{N}_{\mathbb{R}^+}(0, \delta\Sigma), \tag{2.17}$$

to ensure the non-negativity of $\boldsymbol{\phi}$. In this chapter, we chose $\Sigma^{-1} = \boldsymbol{L}^T\boldsymbol{L}$, where $\boldsymbol{L}$ is defined as in (2.13) and the overall amount of smoothness of the solution is governed by the parameter δ (in a similar fashion to λ in the ADMM algorithm). The smaller δ, the smoother the solution. Note that if δ is fixed (which is not the case here), the solution of PIDAL is obtained using MAP estimation.

As shown in Eqn. (2.16), we do not choose a fixed value of δ but assigned to it an inverse-gamma conjugate prior distribution, i.e., $\delta \sim \mathcal{IG}(\alpha_1, \alpha_2)$ with (α_1, α_2) fixed and selected based on WAIC (Watanabe-Akaike Information Criteria) [40]. Since in practice N is large, $f(\phi|\delta)$ dominates $f(\delta)$ (as noted in Chapter 4 of [41]) and the prior distribution $f(\delta)$ has a limited impact on the estimated neutron spectrum. Moreover, as will be shown in the next paragraph, the conjugacy between $f(\boldsymbol{\phi}|\delta)$ and $f(\delta)$ will also simplify the estimation procedure.

2.3.2 Inference

To exploit the posterior distribution $f(\boldsymbol{\phi}, \delta|\boldsymbol{y})$, in this chapter we apply a Markov chain Monte Carlo (MCMC) method which consists of generating random variables distributed according to $f(\boldsymbol{\phi}, \delta|\boldsymbol{y})$. The generated samples are then used to approximate the posterior mean of $\boldsymbol{\phi}$ and associated a posteriori uncertainty intervals. The pseudo-code of the proposed method is summarized in Algo. 1.

The proposed approach is similar to the work in [38] in the sense that we are also using MCMC methods to solve the unfolding problem. However, several important differences can be highlighted. First, as in [38], we estimate the regularization parameters δ, but this is achieved here through a hierarchical Bayesian model (prior distribution assigned to δ) which yields a more computationally efficient algorithm (fewer iterations required) while this parameter is estimated via maximum marginal likelihood estimation in [38]. This approach allows us to also account for the fact that δ is unknown and the additional uncertainty is automatically included when computing confidence regions for $\boldsymbol{\phi}$. Second, here we use a constrained Hamiltonian Monte Carlo methods (as discussed below) which improves the sampler convergence and mixing properties compared to traditional sequential Gibbs updates and random walk-based Metropolis-Hastings updates (as in [38]).

Algorithm 1

HMC unfolding algorithms

Fixed input parameters: $(\alpha_1, \alpha_2), \sigma^2$, number of burn-in iterations N_{bi}, total number of iterations N_{iter}.
Initialization $(k = 0)$
Set $\boldsymbol{\phi}^{(0)} = 1, \delta^{(0)} = \alpha_2/(1 + \alpha_1)$
for $k = 1, \ldots N_{\text{iter}}$ **do**
  Sample $\boldsymbol{\phi}^{(k)} \sim f(\boldsymbol{\phi}|\boldsymbol{y}, \delta^{(k)})$ using HMC
  Sample $\delta^{(k)} \sim f(\delta|\boldsymbol{y}, \boldsymbol{\phi}^{(k)})$ from (2.18)
end for

Set $\widehat{\boldsymbol{\phi}} = 1/(N_{\text{iter}} - N_{\text{bi}}) \sum_{k=N_{\text{bi}}+1}^{N_{\text{iter}}} \boldsymbol{\phi}^{(k)}$

Sampling from $f(\boldsymbol{\phi}, \delta|\boldsymbol{y})$ is achieved by sampling iteratively from $f(\boldsymbol{\phi}|\boldsymbol{y}, \delta)$ and $f(\delta|\boldsymbol{y}, \boldsymbol{\phi})$ (lines 5 and 6 of Algo. 1). It can be easily shown using $f(\delta|\boldsymbol{y}, \boldsymbol{\phi}) \propto f(\boldsymbol{\phi}|\delta) f(\delta)$ that

$$\delta|(\boldsymbol{y}, \boldsymbol{\phi}) \sim \mathcal{IG}\left(\frac{N}{2} + \alpha_1, \frac{\boldsymbol{\phi}^T \Sigma^{-1} \boldsymbol{\phi}}{2} + \alpha_2\right), \tag{2.18}$$

which is straightforward to sample from. The conditional distribution $f(\boldsymbol{\phi}|\boldsymbol{y}, \delta)$ is a non-standard distribution and accept/reject procedures are required to update $\boldsymbol{\phi}$. Due to the potentially large dimensionality of $\boldsymbol{\phi}$ (large number N of bins) and the high correlation between these variables, we resort to a constrained Hamiltonian Monte Carlo (HMC) update which uses the local curvature of the distribution $f(\boldsymbol{\phi}|\boldsymbol{y}, \delta)$ to propose candidates in regions of high probability. This approach allows better mixing properties than more standard random walk alternative strategies. The interested reader is invited to consult [42] for additional details about Hamiltonian Monte Carlo sampling and [43] for an example of application to linear inverse problems involving Poisson noise. The marginal posterior mean $\widehat{\boldsymbol{\phi}}$ is approximated by averaging the generated variables after having removed the first N_{bi} iterations of the sampler which correspond to the burn-in period of the sampler. Similarly, the marginal 95% credible interval for each ϕ_n is computed from the generated samples $\{\phi_n^{(k)}\}_k$. The duration of the transient period N_{bi} and the total number of iterations N_{iter} are set by visual inspection of the chains from preliminary runs. These values are then kept unchanged throughout all the experiments. Note that as mentioned above, by embedding δ in the Bayesian model through $f(\delta)$ and sampling from $f(\boldsymbol{\phi}, \delta|\boldsymbol{y})$, the posterior mean and confidence regions already account for the fact that δ is unknown (they

are computed according to $f(\boldsymbol{\phi}|\boldsymbol{y})$). For completeness, the main parameters of the TiK, PIDAL, and MCMC algorithms are summarized in Table 2.2 below, while the settings used for the three different sources in GRAVEL have been already introduced in Table 2.1.

Table 2.2 Parameters and settings used to unfold the neutron spectra.

Method	Nb. of parameters	Parameters	Value(s)
Tik	1	λ	L-curve [33]
PIDAL	1	λ	user-defined
MCMC	2	(α_1, α_2)	using [40]

2.4 Unfolding Results and Discussion

We assess the performance of proposed algorithm (referred to as MCMC in the remainder of the paper) with GRAVEL [34, 36, 44], Tik (Tikhonov regularization with L-curve method) [33] and PIDAL [29] applied to simulated neutron sources. We consider three sources: 2.5 MeV monoenergetic neutron source, ^{252}Cf and ^{241}AmBe. The data simulation has been performed using the Monte Carlo method detailed in Section 2.2.1 that takes into account the physical process of light output detection with a total number of 5.10^7 detection events, and we use the semi-empirical response matrix described in Section 2.2.1 to unfold the measured light output. In the following experiments, we use the precision matrix $\Sigma^{-1} = \boldsymbol{L}^T\boldsymbol{L}$ as discussed in Section 2.3 for the MCMC algorithm and Tik to be consistent with the PIDAL algorithm. In this paper, we select the optimal (in the sense of the performance measure in Eqn. (2.19)) smoothing parameter of PIDAL based on the ground truth, and the resulting method is denoted as PIDAL-O, which stands for oracle PIDAL, in the sense that this approach uses the value of the smoothing

parameter which gives the best reconstruction performance, which is in practice impossible to obtain without knowing the spectrum to be recovered. This method assumes access to the ground truth spectra, so it can be seen as the optimal MAP estimator and serves as a way to evaluate the difficulty of the unfolding problem.

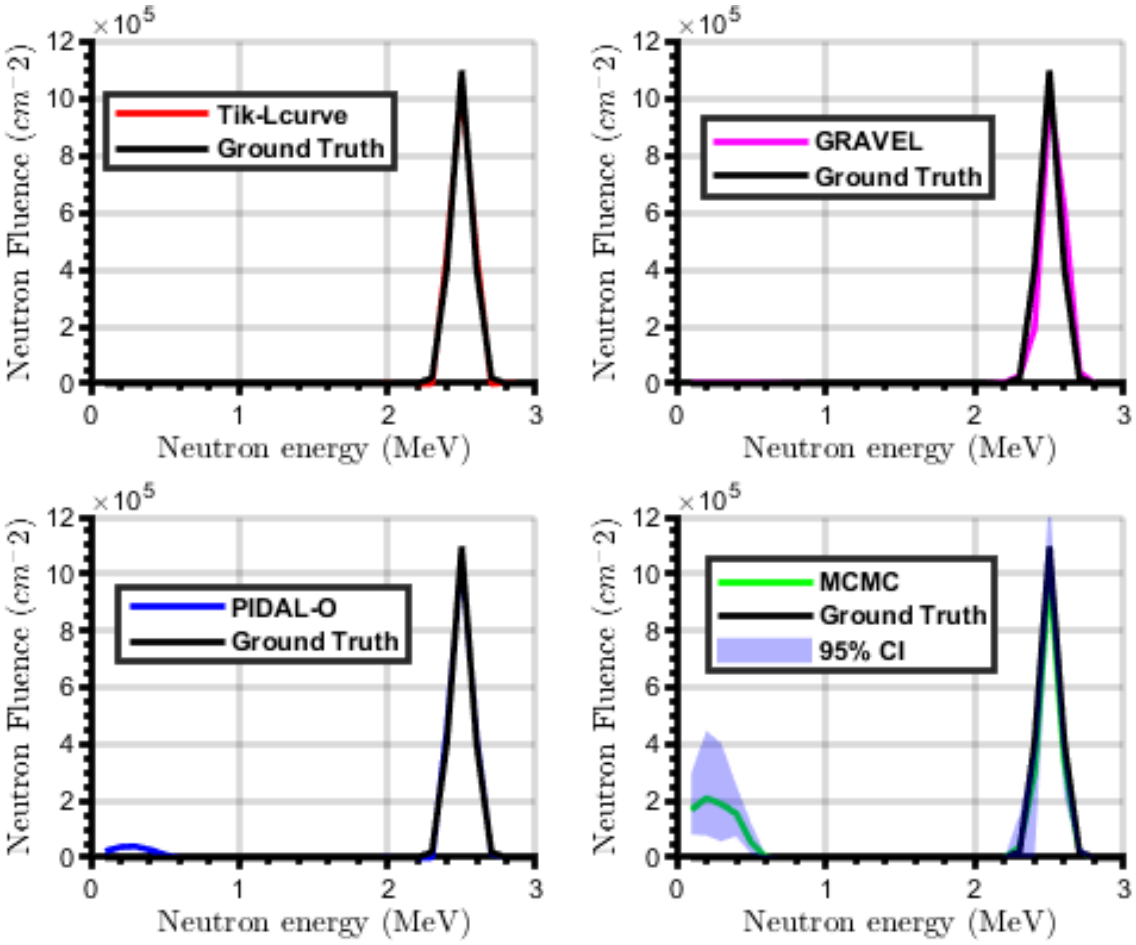


Figure 2.3: Examples of unfolded spectra of the simulated 2.5 MeV monoenergetic neutron source (5.10^7 detection events per light output spectrum). MCMC provides additional uncertainty evaluation through credible intervals (CIs), defined here as the high density regions that contain 95% of the samples drawn from the full posterior distribution (leaving 2.5% on each side). Note PIDAL-O (PIDAL-Oracle) assumes full knowledge about ground truth spectra, so it serves as an estimate of the optimal unfolding algorithm and it is not attainable in actual experimental settings.

Fig. 2.3 shows the unfolded spectra obtained by Tik, GRAVEL, PIDAL-O and MCMC for the simulated 2.5 MeV monoenergetic neutron source. All methods are able to identify the intensity of the peak. MCMC provides additional uncertainty

quantification tools through a posteriori Credible Interval (CI). Here we used a 95% CI corresponding to the high density region that contains 95% of the samples drawn from the full posterior distribution (leaving 2.5% on each side). MCMC identifies a false peak in the lower energy region within which the response matrix is particularly ill-conditioned. This is reflected by the broad posterior confidence region (light blue region) around the posterior mean spectrum. This result is expected since Tik, PIDAL-O and MCMC all impose additional smoothness constraints on the spectrum.

Figs. 2.4 and 2.5 depict the unfolded spectra for the two continuous source (^{252}Cf and ^{241}AmBe). Tik, GRAVEL, PIDAL-O and MCMC all show strong agreement with the ground truth spectrum. In addition, the credible intervals provided by the MCMC algorithm provides additional evidence about regions with higher uncertainty. Fig. 2.6 shows the relative error associated with the unfolded spectra with respect to ground truth for the ^{241}AmBe source. Fig. 2.7 shows the light output obtained as the convolution between the unfolded spectra and the response matrix compared to the ground truth light output. The four methods show very good agreement with the ground truth. This result illustrates one of the main challenges of the neutron unfolding problem, where several different unfolded spectra can lead to similar fits to the data to be deconvolved. Note that the relative error plots and generated light output plots for ^{252}Cf lead to the same conclusions as those presented using ^{241}AmBe, thus they are omitted here to reduce redundancy.

We use the Spectral Angle Mapper (SAM) [45] between the unfolded spectrum ($\hat{\boldsymbol{\phi}}$) and the known ground truth ($\boldsymbol{\phi}$) to quantify the unfolding performance of

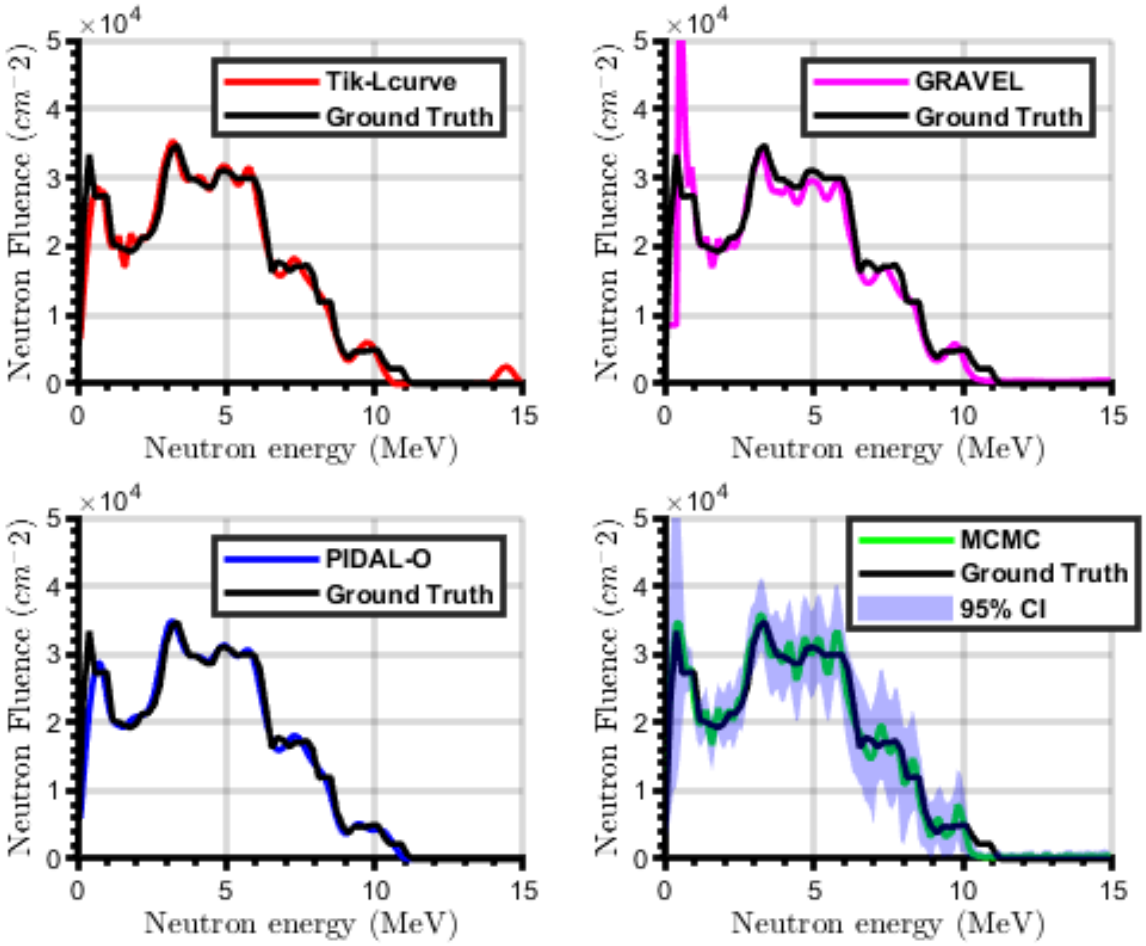


Figure 2.4: Examples of unfolded spectra of the simulated ^{241}AmBe neutron source (5.10^7 detection events per light output spectrum).

the different methods. Because the ground truth neutron spectra and response matrix have different neutron energy resolutions, we adopted SAM as opposed to standard Mean Square Error (MSE) as SAM is scale-invariant. Indeed, the SAM criterion relies on the spectral angle between $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$, which is small when $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ present similar shapes. As a result, similar spectra lead to values of SAM close to 0. The energy bounds listed in Table 1 were applied to the GRAVEL unfolded spectra to calculate the SAM.

$$SAM(\boldsymbol{\phi}, \hat{\boldsymbol{\phi}}) = \arccos\left(\frac{\boldsymbol{\phi}^T \hat{\boldsymbol{\phi}}}{||\boldsymbol{\phi}||_2 ||\hat{\boldsymbol{\phi}}||_2}\right). \tag{2.19}$$

Table 2.3 summarizes all the SAMs which appear to be in agreement with the

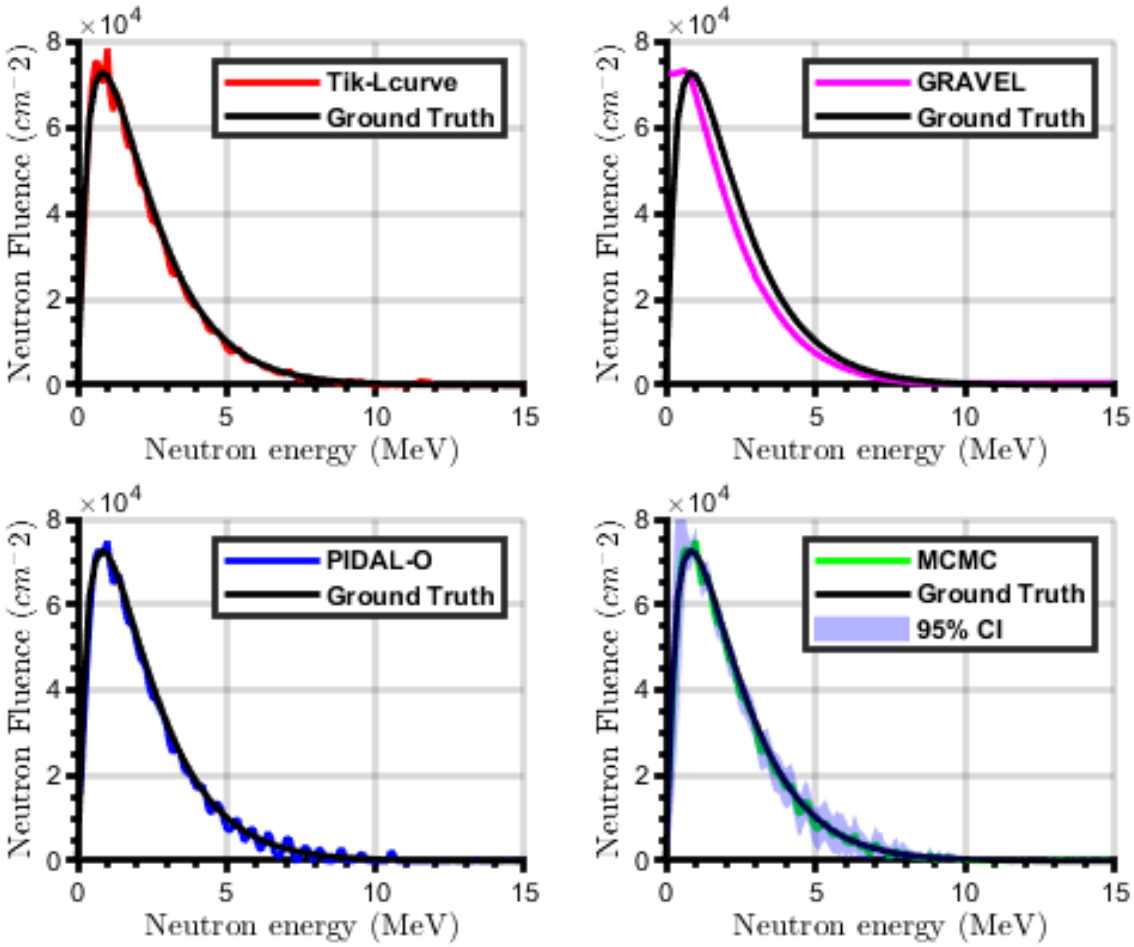


Figure 2.5: Examples of unfolded spectra of the simulated ^{252}Cf neutron source (5.10^7 detection events per light output spectrum).

qualitative results as shown in Figs. 2.3 to 2.5. Notably, MCMC, PIDAL and Tik all provided the competitive results based on SAM for the two continuous source, but MCMC automatically estimates the amount of regularization required from the data with additional credible interval.

Table 2.3 Spectral Angle Mapper (degrees) obtained using the different unfolding methods for the three sources (5.10^7 detection events per light output spectrum). Note PIDAL-O (PIDAL-Oracle) assumes full knowledge about ground truth spectra, so it serves as an estimate of the difficulty of the unfolding problem and it is not attainable in actual experimental settings

Method / Neutron Source	Tik	GRAVEL	PIDAL-O	MCMC
DD	3.54	14.23	3.97	18.75
$^{241}AmBe$	6.26	4.6 13.30	6.29	5.13
^{252}Cf	2.97	4.73 14.14	3.47	2.69

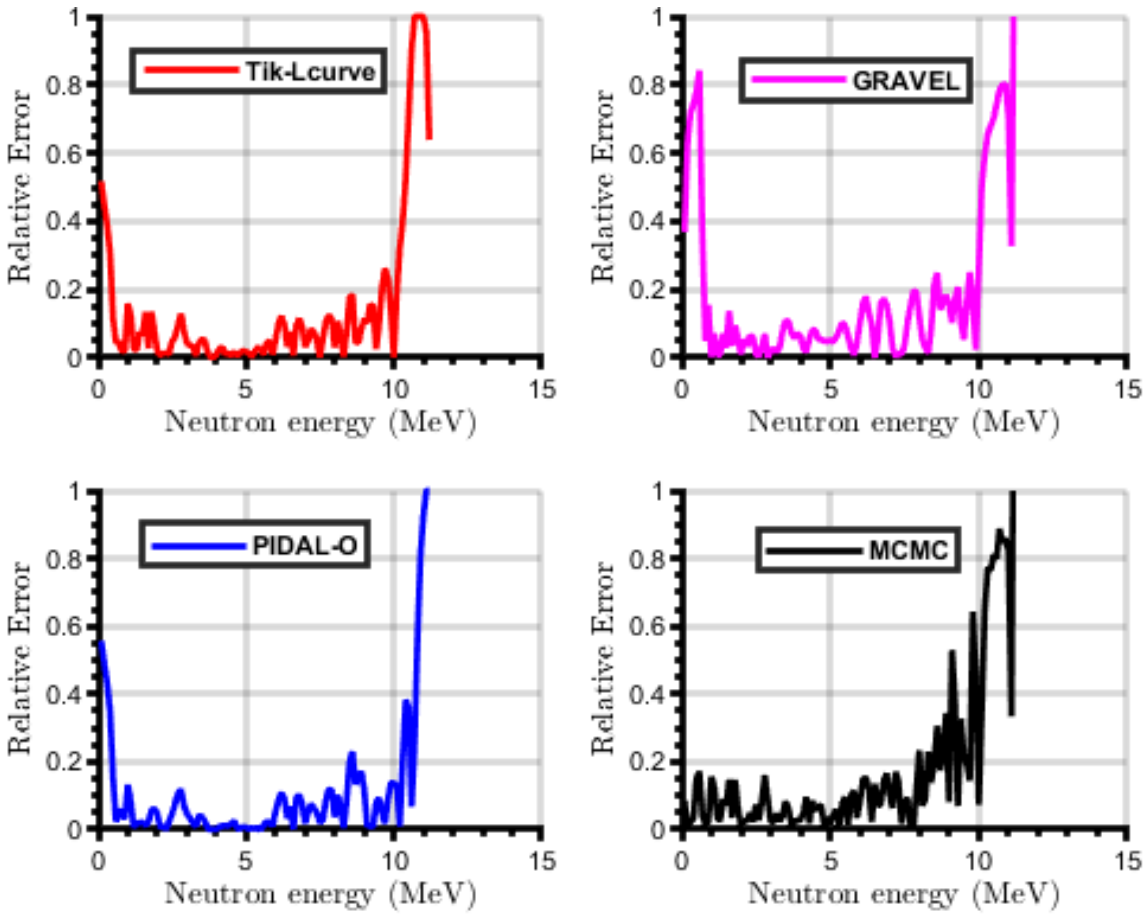


Figure 2.6: Relative error plots of unfolded spectra of the simulated ^{241}AmBe neutron source (5.10^7 detection events per light output spectrum) with respect to the Ground truth. Note PIDAL-O (PIDAL-Oracle) assumes full knowledge about ground truth spectra, so it serves as an estimate of the optimal unfolding algorithm and it is not attainable in actual experimental settings.

In safeguards, security, and non-proliferation applications, it is often realistic to have a weak neutron signal that can be overwhelmed by an intense gamma-ray background [46]. Therefore, it is of considerable interest to examine the robustness of the algorithms as the number of detection event decreases (weak source and/or short integration time). We assess the robustness of the different algorithms using simulated data of ^{252}Cf and ^{241}AmBe, for event counts ranging from 5×10^2 up to 5×10^6. Note that for the most challenging scenarios, e.g., using only 5×10^2 total counts across the $M = 600$ light output bins, the average counts per bin fall below 1 for both ^{252}Cf and ^{241}AmBe, with 480 empty bins on average for ^{241}AmBe

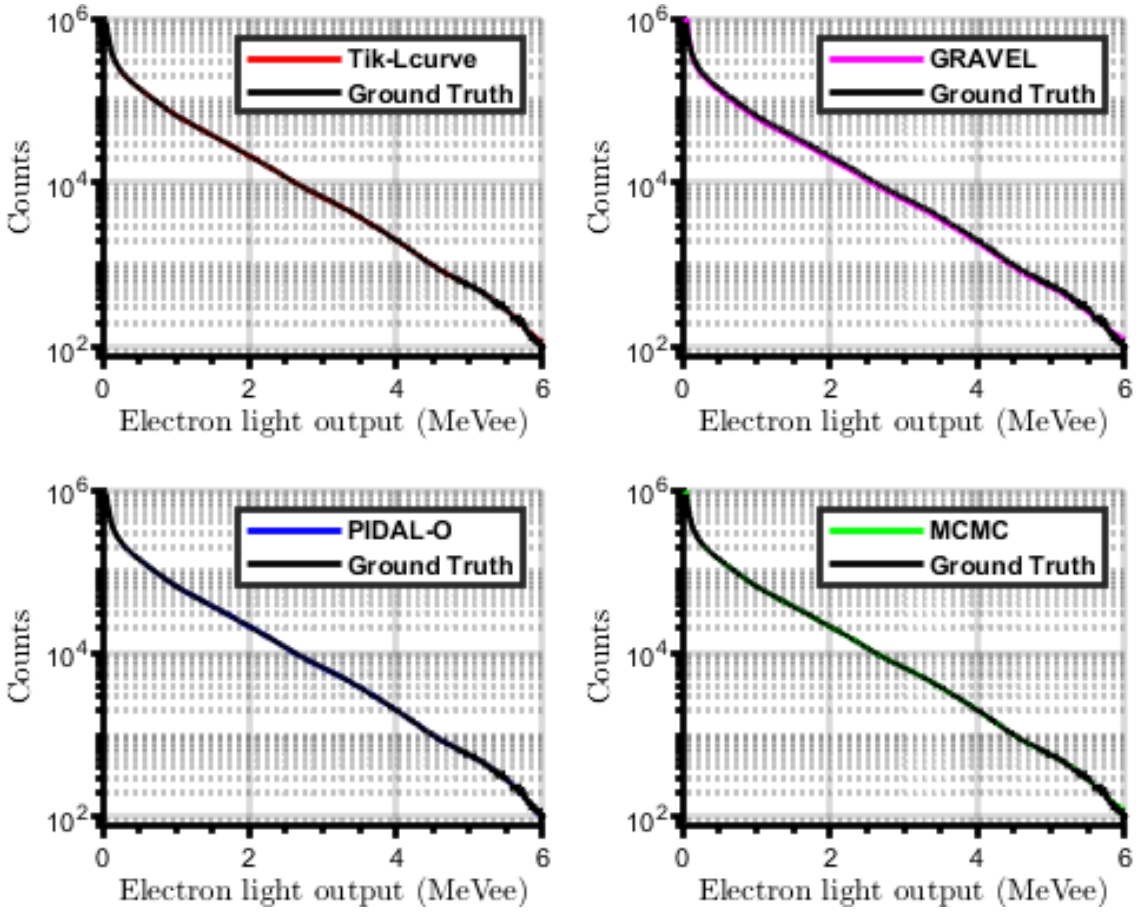


Figure 2.7: Examples of light output spectra generated using the unfolded spectra of the simulated ^{241}AmBe neutron source (5.10^7 detection events per light output spectrum) compared with ground truth light output. Note PIDAL-O (PIDAL-Oracle) assumes full knowledge about ground truth spectra, so it serves as an estimate of the optimal unfolding algorithm and it is not attainable in actual experimental settings

and 520 empty bins for ^{252}Cf. This further motivates the use of the Poisson noise model in our unfolding procedure. The results are summarized in Fig. 2.6 and Table 2.4. Note that GRAVEL failed to converge for both sources at numbers of counts lower than 5×10^4, which is denoted as N/A.

As mentioned in Section 2.3, PIDAL can be seen as a special case of the proposed hierarchical model where the hyperparameter δ is fixed as opposed to random. With appropriately tuned regularization parameters, Tik, PIDAL and MCMC demonstrated the competitive robustness against low counts. However,

the proposed MCMC algorithm automatically adjusts this parameter and does not require exact knowledge about the ground truth.

Table 2.4 Unfolding performance (average SAM, in degree) as a function of the total number of detection event (best result per row in bold). Values in brackets represent standard deviations computed over 50 Monte Carlo realizations. Note PIDAL-O (PIDAL-Oracle) assumes full knowledge about ground truth spectra, so it serves as an estimate to the difficulty of the unfolding problem and it is not attainable in actual experimental settings.

Neutron Source	Counts	Tik	GRAVEL	PIDAL-O	MCMC
$^{241}AmBe$	5×10^6	8.99 (1.96)	14.71 (2.99)	**7.47** (0.77)	7.99 (0.29)
	5×10^5	9.87 (0.46)	15.81 (1.90)	**8.93** (0.86)	9.89 (0.40)
	5×10^4	11.84 (0.49)	N/A	**10.96** (1.24)	12.79 (0.65)
	5×10^3	15.25 (0.62)	N/A	**14.64** (1.54)	17.06 (1.11)
	5×10^2	19.40 (3.75)	N/A	**17.18** (1.41)	22.04 (2.61)
^{252}Cf	5×10^6	4.69 (0.45)	14.59 (1.02)	4.54 (0.60)	**4.28** (1.12)
	5×10^5	5.05 (0.84)	15.51 (1.60)	5.78 (0.75)	**4.62** (1.06)
	5×10^4	7.06 (1.11)	N/A	7.20 (1.02)	**6.33** (1.68)
	5×10^3	12.25 (1.14)	N/A	10.35 (2.03)	**10.01** (2.26)
	5×10^2	16.97 (2.51)	N/A	**14.57** (3.22)	22.73 (1.96)

In practical applications, systematic errors in the unfolded spectra may arise because of an inaccurate calibration of the detector or a drift in the operating conditions, e.g. temperature. In such cases, the presented methods are expected to exhibit a similar energy bias in the reconstructed spectrum since no strong prior information is incorporated into the algorithms. The unfolding of a known monoenergetic spectrum, e.g., from ^{137}Cs, with suitable gamma-ray response matrix, could be used to mitigate and correct for such systematic errors. We implemented the Tik, PIDAL-O and the proposed MCMC unfolding algorithm in Matlab R2017b on an 2GHZ Intel processor with 6GB of RAM. The maximum number of iteration for Tik and PIDAL are fixed at 24000 but the algorithms generally converge and are stopped well before this number of iterations. Within the MCMC algorithm, we generated sequentially 24000 samples (after the burn-in period of the sampler) for all the simulation results presented in this paper. Tik

and PIDAL-O calls Tik and PIDAL to search for the best smoothing parameter. The tuning of hyperparameters of MCMC algorithm is done using WAIC (Watanabe-Akaike Information Criteria) [40]. We used the compiled version of GRAVEL available through RSICC (UMG package version 3.3). The average run time of the algorithms to analyze one spectrum is presented in Table 2.5. As shown in Table 2.5, the enhanced unfolding performance of the MCMC method comes with a significantly higher computational cost than Tik, GRAVEL and PIDAL (for a fixed value of the smoothing parameter) because the sequential nature of the sampler and the number of iterations required to estimate the posterior mean and credible intervals. Different choices of parameters for MCMC results in the significant discrepancy of run time for ^{241}AmBe and ^{252}Cf. In actual experiment, Tik (with L-curve Method) and PIDAL-O are called 70 times to perform a log scale search to find the best smoothing parameter prior a full run, while MCMC are called 6 times to perform a log scale search. However, it is worth noting that the hyperparameter selection procedure and the algorithm implemented has not been optimized for fast analysis, and it is possible to accelerate the method using C/C++ implementations.

Table 2.5 Average computational time to analyze one spectrum (in seconds) over 100 runs. Note all the reported time here excludes the additional parameter tuning time cost.

Method / Neutron Source	Tik	GRAVEL	PIDAL	MCMC
$^{241}AmBe$	0.38	900	0.45	83.39
^{252}Cf	0.71	60	0.53	40.42

2.5 Conclusions

We have proposed a hierarchical Bayesian approach to solve the neutron spectrum unfolding problem, which differs from previous work [26, 38] by using an efficient constrained Hamiltonian Monte Carlo method and a hyper-prior on the hyper-parameter. The new MCMC algorithm shows improvement in performance compared to traditional approaches, such as Tik [33], GRAVEL [34, 47, 44] and PIDAL [29] on simulated data (^{252}Cf and ^{241}AmBe) in terms of accuracy with additional uncertainty evaluation through credible interval. This chapter further demonstrates the potential benefits of Bayesian methods for solving unfolding problems, because they provide a formalized manner in which to integrate existing prior knowledge within the estimation procedure. In this chapter, we have focused on synthetic data generated from reference neutron spectra and a known response matrix (ground truth available). In future work, the performance of the algorithm will be evaluated using measured data (simulated and measured response matrices) for organic scintillators. Efforts should in particular concentrate on robustness of the methods with respect to detector imperfections and background/spurious detections. Additional types of detectors with spectroscopic capability, e.g., Bonner sphere spectrometers, silicon telescopes, and superheated emulsions will also be investigated. The present unfolding method could also be coupled to classification algorithms to infer the type and amount of fissile material in unknown neutron sources, for nonproliferation and safeguarding applications. Approximate Bayesian methods will also be investigated for robust unfolding with reduced processing burden.

CHAPTER III

A Graphical Model for Fusing Diverse Microbiome Data

3.1 Introduction

In this chapter we introduce a Bayesian graphical model for joint modeling and fusing high dimensional count data collected from different sensors with no explicit correspondences between their feature sets. Our model is relevant to the many areas of multi-modality fusion where data is collected from diverse but incommensurate sensor modalities. Examples include multi-view learning in computer vision and automated language translation in natural language processing. However, this paper focuses on a particularly timely application: the fusion of microbiome data from diverse microbial communities.

Microbiomes exist in diverse environments and are critical to sustaining life, balancing ecosystems, and producing antibiotics, among many other functions. Microbiomes consist of communities of microbes that interact with each other to maintain stability and resilience to environmental conditions and microbial intrusions from competitors. It has therefore been of great scientific interest to quantify changes in microbiome communities due to changing conditions using experimental data. For example, one area of study is the rhizosphere, which is a community of microbial species living around plant root systems, known to be

sensitive to environmental factors [48]. Another area of study is the spectrum of responses of microbiomes to stressors, collectively called the microbial exposome [49].

One of the principal sensing platforms used to study microbiome communities applies gene sequencing to a microbiome sample, e.g., collected from the gut, the soil, or other environments. A common way to obtain a global profile of a microbial community is to perform gene sequencing on a biological sample. For example, RNA-Seq measures gene expression in a community by quantifying the number of times each gene transcript occurs in the pool of sequenced RNAs. Each microbial species in the community is represented by its own unique set of transcripts, i.e., its transcriptome, and fusing information from different transcriptomes yields the global profile of gene expression across all species in the community. This type of analysis is known as metatranscriptomics and it provides a functional profile of the community that can complement the gene taxonomic profiling provided by metagenomics [50, 51, 52]. The resulting datasets consists of species abundance (count of RNA occurrences) for different samples obtained from various communities. This paper introduces a Bayesian graphical model for the metatranscriptomics problem, and inference is performed using a scalable variational EM inference method. Notably, our model can capture patterns of similarity between histograms of different species' gene expression without inter-species genome-to-genome mappings or knowledge of inter-species transcriptomic pathway correspondences.

The main feature of our model is that it estimates the global covariance structure of gene expression when the observations are in the form of count vectors

produced by RNA-SEqn. Correlations between transcript abundances are informative about the effect of environmental conditions on microbial communities [53]. In particular, the global covariance matrix captures inter- and intra-species interactions. For example, the expression of a single gene in a species can influence other gene expressions in that species or the gene expressions of other community members. We propose a latent variable graphical model that can capture the hidden factors underlying such dependencies.

The main assumption underlying our proposed model is the existence of a hidden low-dimensional continuous latent space that can explain the observed data. We model the observations as conditionally multinomial distributed given the latent variables, which are assumed to be multivariate Gaussian with a low-rank covariance structure. Due to the lack of conjugacy between the Gaussian and multinomial distributions, exact Bayes inference is not tractable. We, therefore, adopt a Bayes variational inference approach [54, 55] to develop an algorithm for estimating the parameters of the proposed model and projecting the data to the latent space.

The proposed model can be contrasted with previously introduced latent variable models used in multi-view learning and dimensionality reduction. Factor analysis (FA) [56] is a classical method that is a generalization of Principal Component Analysis (PCA) [57] and Probabilistic PCA [58]. FA decomposes the observed data matrix into a low-dimensional set of factor loadings and factor scores, imposing a low-rank constraint on the covariance matrix. Like our proposed model, the FA model also assumes a low-dimensional Gaussian latent space but it does not account for the counting nature of the observed data.

Several latent variable models have been proposed for counting observations. These include Latent Semantic Analysis [59], Multinomial PCA [60], and Latent Dirichlet Allocation (LDA) [61]. LDA is the most closely related model to the model proposed here since it is also a Bayesian graphical model for count data and uses multinomial distribution. The main difference is that LDA uses a Dirichlet-distributed latent space instead of a Gaussian-distributed latent space. Our Gaussian distributed latent space makes it possible to recover a non-trivial covariance structure among the count variables, unlike LDA [62, 63].

Another way to capture the covariance structure of the observed variables is to ignore the counting nature of the data and use Gaussian Markov random fields (GMRF) [64] to directly estimate the covariance, or Gaussian Graphical Models (GGM) [65] to enforce sparsity on the inverse of the covariance estimate. There have been extensions of the GGM to handle multinomial observations using copulas [66] that have been applied to microbiome analysis [67, 68, 69]. There is also an ongoing effort to extend the GGM to the multiple datasets settings where there is an assumed common precision matrix across the datasets [70, 71, 72, 73, 74, 75, 76, 77]. Notably, [78] extends previous optimization-based approaches to the hierarchical Bayesian setting along with a scalable and efficient inference method. However, this line of work assumes a common feature space across multiple datasets, whereas in our case the features are distinct for different microbial communities.

In the field of computational ecology, there has been a related line of work on joint species distribution modeling (JSDM) [79, 80, 81] to model multiple related abundance datasets. The proposed work differs from JSDM mainly in how we

represent the environmental covariates. In JSDM, the environmental covariates are used to infer the species abundance through the generalized linear model, whereas in this chapter we explicitly represent the covariates through latent variables. With the latter more suitable for applications where the covariates are discrete descriptors of environments such as the binary case (the presence of a bacterium that produces koreenceine antibiotics) we are considering in 3.3.2.

Inference in latent variable models, like the one we propose here, can be challenging. This is especially difficult when there is a lack of conjugacy between the distributions of the latent variables and the observed variables. One approach is to perform point estimation for both the latent variables and the parameters in an alternating fashion [82], but this is prone to over-fitting [83] and convergence issues. Another approach is to use Markov Chain Monte Carlo (MCMC) methods, which can be computationally expensive [84], especially in high dimensions. As an alternative, variational Bayes inference has shown much promise [55]. Note that Variational Bayes is not a general purpose method and must be tailored to the specific statistical model [85]. When there is a lack of conjugacy, as is the case for the multinomial-Gaussian model in this paper, local variational bound approximations are often adopted [54]. Additionally, when there is a problematic expression in the joint density, such as the $LogSumExp$ or $LogGamma$ function, which may prevent the inference of the latent variables, surrogate optimization transfer based on Taylor series expansion can be applied to approximate the non-linear function either with linear [62] or quadratic [86, 87, 88, 89, 90] functions. We adopt such a local variational bound approach for deriving an inference algorithm for our proposed model.

The proposed model has connections with multi-view learning, text embedding methods, and manifold learning. Supervised PCA [91, 92], Partial Least Squares [93], Canonical Correlation Analysis [94], and Multimodal Factor Analysis (MMFA) [95] allow fusing multi-view data into a common low-dimensional latent space. Among them, only MMFA is applicable to non-Gaussian observations, which, however, does not apply to vectors of count data with observed covariance. Furthermore, the MMFA assumes a non-random latent space, which is known to be prone to over-fitting [83]. Variational auto-encoder-based deep neural network models [96, 97] are often implemented with only a single latent variable to explain multiple modalities. Such autoencoders are implemented by maximizing evidence lower bound (ELBO) that exploits the product of experts framework to combine multiple modalities. [98, 99] use an equalized mixture of experts to combine modality-specific encoder predictions. [100] separates the latent variables as joint and individual where joint latent variables are common for each input modality and individual latent variables are only used to generate the corresponding observations. Deep generative models have shown recent promise for modeling densities of complex structured data, providing accurate predictions for out-of-sample inputs when the number of training samples is large. However, most microbiome datasets, which are the focus of this paper, have few samples, often many fewer than the number of features. Thus deep models are prone to overfitting such datasets. Furthermore, unlike the proposed model, there is no straightforward way to predict the covariance structure of the observation space using deep learning models.

Count vector data also arises in natural language processing (NLP), where a

sentence or a document can be described using a bag-of-words representation. Early NLP models, such as Latent Semantic Analysis/Indexing [101], perform factorization of the count matrix using Singular Value Decomposition, but do not account for the multinomial nature of the data. More recent algorithms, such as Word2Vec [102] and Glove [103], model the sequence of words using a context window. Contemporary contextual word embedding methods [104], on the other hand, employ deep-learning models. ELMO [105] and BERT [106] are among the most popular, which exploit Recurrent Neural Networks and Transformers, to model the hidden dynamics between consecutive words, respectively. Note that, many NLP algorithmsuse Markovian dynamical models for dependence between consecutive words. However, gene indices for microbiome assays are not ordered making NLP inapplicable.

While our proposed model uses dimension-reducing latent variable parameterization, it differs significantly from manifold learning algorithms. Such algorithms learn low-dimensional representations using methods such as Multidimensional Scaling [107], Kernel PCA [108], Isomap [109], Local Linear Embedding [110], and t-distributed Stochastic Neighbor Embedding [111]. Unlike the Bayesian hierarchical models introduced by our model, manifold learning algorithms are not capable of specifying the predictive distribution of the latent variables, performing model integration over different feature categories (species), or specifying a covariance model.

We summarize our contributions as follows. First, we propose a novel multinomial-Gaussian graphical model to fuse and capture the low-rank covariance structure in counting data of disparate types. Our low-dimensional continuous latent space

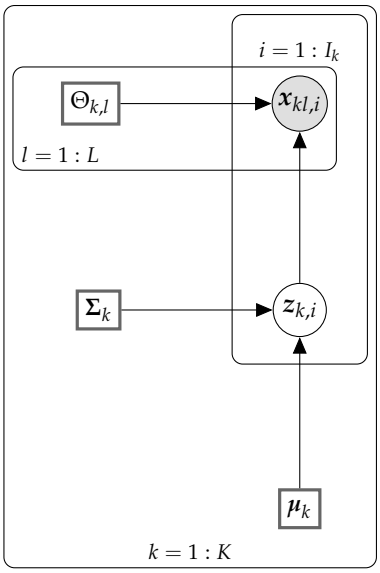


Figure 3.1: Graphical model representation of the proposed latent variable model. $x_{kl,i}$ corresponds to the ith sample of community l collected from environment k. The variables $\{x_{kl,i}\}_{l=1:L}$ share a common low-dimensional latent variable $z_{k,i}$ that captures the hidden causes of the observations.

formulation provides dimensionality reduction that can be used for visualization of the count vectors on a common space.

Second, we develop a novel and computationally scalable optimization algorithm based on variational inference to fit the proposed model, which exploits variational local bound approximations. Third, we validate and illustrate the model and its inference algorithm on a synthetic dataset and a real-world bacterial microbiome dataset.

3.2 Proposed Model

In this section, we formally define our proposed model and its corresponding variational inference algorithm. Lastly, we discuss computational complexity.

3.2.1 Notation

We denote the ith count vector replicate for the lth species as $\boldsymbol{x}_{kl,i} \in \mathbb{Z}_+^{d_l}$, where k indexes the experimental condition, and d_l denotes the total number

of transcripts for species l. The total number of experimental conditions from which the samples are collected is denoted as K, and the total number of species in the model community is denoted as L, hence $l = 1, \ldots, L$ and $k =, \ldots, K$. For each experimental condition, different numbers of identically distributed samples are collected. Hence, we denote the total number of samples for the experimental condition k as I_k. Concisely , the dataset for experimental condition k is $D_k = \{\{\boldsymbol{x}_{kl,i}\}_{l=1}^{L}\}_{i=1}^{I_k}$.

3.2.2 Latent Variable Model

We model the observed multi-species model community data as generated from a low-dimensional latent variable generative model. Under this model, the data are conditionally multinomial distributed given the latent variables, which are themselves Gaussian distributed with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma_k}$. As will be shown below, the model fuses the observed data across species and it induces a low-rank decomposition of the population transcriptome covariance. Let $\boldsymbol{z}_{k,i} \in \mathbb{R}^{d_z}$ be the latent variable assigned for the data sample $D_{k,i}$. $\boldsymbol{z}_{k,i}$ thus has the following multivariate normal prior distribution:

$$p(\boldsymbol{z}_{k,i}; \boldsymbol{\mu}_k, \Sigma_k) = \mathcal{N}(\boldsymbol{z}_{k,i}; \boldsymbol{\mu}_k, \Sigma_k), \tag{3.1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{d_z}$ is the prior mean vector and $\Sigma_k \in \mathbb{S}_{++}^{d_z}$ is the positive definite prior covariance matrix. The observed data consists of count vectors of the transcriptomes, which are modeled as multinomial distributed [63]. We model the conditional distributions of the observed count vectors of species l as follows:

$$p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) = Mu(\boldsymbol{x}_{kl,i}; N_{kl,i}, \mathcal{S}(\Theta_{kl}\boldsymbol{z}_{k,i})), \tag{3.2}$$

where $N_{kl,i}$ is the total number of counts of the ith data sample of species l and Mu denotes the multinomial distribution with the form, $Mu(\boldsymbol{x}; N, \boldsymbol{p}) = \frac{N!}{x_1!x_2!\ldots x_D!}\prod_{d=1}^{D} p_d^{x_d}$. Note that we have introduced one more model parameter for each species, specifically $\Theta_{kl} \in \mathbb{R}^{d_l \times d_z}$, that maps lower dimensional latent space to the higher dimensional observation space of species l. Also note that both the latent variable $\boldsymbol{z}_{k,i}$ and the parameter Θ_{kl} are real-valued. Therefore, to provide a proper simplex support set for the multinomial distribution, we use the soft-max function, $\mathcal{S}(\boldsymbol{\eta})_d = \exp \eta_d / \sum_{d'=1}^{D} \exp \eta_{d'}$, where $\mathcal{S}(\boldsymbol{\eta})_d$ is the dth element of probability vector $\mathcal{S}(\boldsymbol{\eta})$ and $\eta = \Theta_{kl}\boldsymbol{z}_{k,i}$ for notational simplicity. The output of this function is a proper probability vector, i.e., $\sum_{d=1}^{D} \mathcal{S}(\boldsymbol{\eta})_d = 1$ and $\mathcal{S}(\boldsymbol{\eta})_d \geq 0$ for all $d = 1, \ldots, D$. See Fig. 3.1 for a graphical representation of the proposed model.

The lower dimension of the latent variables is a key feature of our model since it explicitly induces lower rank constraints on the observation covariance matrix, as will be explained in Section II-D, leading to a reduction in the total number of model parameters. It also improves the computational efficiency of the optimization algorithms, as shown in Section II-E. A theoretical justification is supplied by the manifold hypothesis [112], which holds that most naturally occurring signals lie in a lower dimensional space, in addition to the principle of Occam's razor [113], which holds that choosing less complex models leads to better and more stable performance.

Although it is natural to model the observed counts as multinomial distributed, it may not be obvious why we use Gaussian latent variables for the latent space.

A conjugate distribution such as Dirichlet may seem more natural than the Gaussian distribution, which is not conjugate to Multinomial. However, the components of the Dirichlet distribution are nearly independent [62], hence it is non-trivial to capture the correlations between the hidden components. On the other hand, the Multivariate normal distribution has a covariance parameter that specifically captures the correlation between the hidden components. This is useful for modeling the correlation between multiple datasets. Similar model assumptions are also adopted in topic models [62, 114], categorical PCA [89], and Gaussian process classification [90]. Note that, although the communities are dependent through the latent variables, the experimental conditions are modeled as independent. Hence, there is no coupling between the experimental conditions and thus we fit independent models for each condition.

The joint log-likelihood of the proposed model is of the form $\sum_{k=1}^{K}\sum_{i=1}^{I_k}\log p(\boldsymbol{z}_{k,i}, D_{k,i})$, where:

$$\begin{aligned}\log p(\boldsymbol{z}_{k,i}, D_{k,i}) &= \log p(\boldsymbol{z}_{k,i}) + \sum_{l=1}^{L}\log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) \\ &= -\frac{1}{2}\left[(\boldsymbol{z}_{k,i}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{z}_{k,i}-\boldsymbol{\mu}_k) + \log|\Sigma_k|\right] \\ &\quad + \sum_{l=1}^{L}\sum_{d=1}^{D}\boldsymbol{x}_{kl,id}(\Theta_{kl,d}\boldsymbol{z}_{k,i} - \mathrm{lse}(\Theta_{kl}\boldsymbol{z}_{k,i})) + const,\end{aligned} \tag{3.3}$$

in which lse denotes the log-sum-exp function, i.e., log of the denominator of the soft-max function, $\mathrm{lse}(\boldsymbol{\eta}) = \log\sum_{d=1}^{D}\exp\eta_d$, and we suppress the deterministic parameters to avoid clutter. Taking the expectation with respect to $\boldsymbol{z}_{k,i}$ is tractable for the linear and quadratic terms, but intractable for the lse term. We describe an asymptotic approximation in the next section.

3.2.3 Optimization

Next, we develop a variational EM maximum likelihood algorithm [54, 55] to infer the deterministic parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and Θ_{kl}. The main objective is to maximize the likelihood of the observations under the model. The algorithm comprises two alternating steps: i) the Expectation step (E-step), where we integrate out the latent variables, ii) the Maximization step (M-step), where we optimize the model parameters to maximize the marginal likelihood.

Objective

The proposed model uses Gaussian latent variables for the multinomial observations. Due to the lack of conjugacy between Gaussian and Multinomial distributions, the likelihood function is not closed form. Specifically, integrating out the latent variables becomes intractable (See Section 3.2.3 for the details). Hence, we resort to variational inference, in which a lower bound on the likelihood function is derived and maximized. This lower bound is obtained by approximating the posterior distributions of the latent variables. In variational inference, the objective is to minimize the distance (KL-divergence) between the approximate and exact posterior distributions. This objective can be expressed for a single latent variable $z_{k,i}$ as follows:

$$
\begin{aligned}
\mathbb{KL}(q_{\lambda_{k,i}}|p) &= \mathbb{E}_{q_{\lambda_{k,i}}} \log \big[\frac{q(\boldsymbol{z}_{k,i};\boldsymbol{\lambda}_{k,i})}{p(\boldsymbol{z}_{k,i}|D_{k,i})}\big] \\
&= \mathbb{E}_{q_{\lambda_{k,i}}} \log \big[\frac{q(\boldsymbol{z}_{k,i};\boldsymbol{\lambda}_{k,i})}{p(\boldsymbol{z}_{k,i},D_{k,i})} p(D_{k,i})\big] \\
&= \mathbb{E}_{q_{\lambda_{k,i}}} \big[\log q(\boldsymbol{z}_{k,i};\boldsymbol{\lambda}_{k,i}) - \log p(\boldsymbol{z}_{k,i},D_{k,i})\big] \\
&\quad + \log p(D_{k,i}),
\end{aligned}
\tag{3.4}
$$

where $\boldsymbol{\lambda}_{k,i}$ corresponds to the set of parameters of the approximate posterior distribution $q(\boldsymbol{z}_{k,i}; \boldsymbol{\lambda}_{k,i})$. The expectation operator is defined as $\mathbb{E}_{q_\lambda} f(z) = \int f(z) q(\boldsymbol{z}; \boldsymbol{\lambda}) dz$. Note that the evidence (marginal likelihood) $p(D_{k,i})$ does not depend on $\boldsymbol{z}_{k,i}$. Hence, the negative of the expectation term forms a lower bound on the log evidence since the KL distance is always non-negative. This function is known as evidence lower bound (ELBO) and it is the objective function that is maximized in variational EM. The ELBO has the following form:

$$\mathcal{L} = \sum_{i=1}^{I} \sum_{k=1}^{K} \mathbb{E}_{q_{\lambda_{k,i}}} \big[\log p(\boldsymbol{z}_{k,i}, D_{k,i}) - \log q(\boldsymbol{z}_{k,i}; \boldsymbol{\lambda}_{k,i}) \big], \tag{3.5}$$

where the first term in the expectation corresponds to the joint distribution of the latent variable $\boldsymbol{z}_{k,i}$ and the associated observed data $D_{k,i}$. The second term corresponds to the log of the approximate posterior distribution. The joint distribution has the following form:

$$\log p(\boldsymbol{z}_{k,i}, D_{k,i}) = \log p(\boldsymbol{z}_{k,i}) + \sum_{l=1}^{L} \log p(\boldsymbol{x}_{kl,i} | \boldsymbol{z}_{k,i}). \tag{3.6}$$

The expressions for $p(\boldsymbol{x}_{kl,i} | \boldsymbol{z}_{k,i})$ and $p(\boldsymbol{x}_{kl,i} | \boldsymbol{z}_{k,i})$ are given in Eqn. 3.2 and Eqn. 3.1, respectively. We approximate the posterior distribution of $\boldsymbol{z}_{k,i}$ as Gaussian with the following form:

$$q(\boldsymbol{z}_{k,i}; \boldsymbol{\lambda}_{k,i}) = \mathcal{N}(\boldsymbol{z}_{k,i}; \boldsymbol{m}_{k,i}, \boldsymbol{S}_{k,i}), \tag{3.7}$$

where $\boldsymbol{\lambda}_{k,i} = \{\boldsymbol{m}_{k,i}, \boldsymbol{S}_{k,i}\}$ is the set of free parameters. Specifically, $\boldsymbol{m}_{k,i}$ is the posterior mean and $\boldsymbol{S}_{k,i}$ is the posterior covariance. The expectation of the approximate posterior distribution in Eqn. 3.5 corresponds to the Gaussian entropy function, which has a closed-form expression. However, the expectation of the joint distribution is intractable to compute. Next, we present an approximation to resolve the issue.

An upper bound on the LSE

To see why the conditional expectation is intractable, note that the explicit form of the log-likelihood of $\boldsymbol{x}_{kl,i}$ is a multinomial distribution:

$$\log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) = \sum_{d=1}^{D} \boldsymbol{x}_{kl,id}(\Theta_{kl,d}\boldsymbol{z}_{k,i} - \text{lse}(\Theta_{kl}\boldsymbol{z}_{k,i})). \tag{3.8}$$

Taking expectation corresponds to integrating out Gaussian distributed $\boldsymbol{z}_{k,i}$. The conditional expectation of the first term is easily determined since it linearly depends on $\boldsymbol{z}_{k,i}$. However, the expectation of the second term, which requires integrating $\boldsymbol{z}_{k,i}$ over the lse function, is intractable to compute in a closed form. To overcome this issue, we perform quadratic surrogate optimization transfer (See Appendix 3.5.2), in which a quadratic approximation to the lse function [86] is applied. This results in an upper bound on the multinomial log-likelihood. This approximation uses the second-order Taylor series expansion with a fixed Hessian matrix. Particularly, the quadratic upper bound takes the following form (See Appendix 3.5.3 for more details):

$$\text{lse}(\Theta_{kl}\boldsymbol{z}_{k,i}) \leq \frac{1}{2}\boldsymbol{z}_{k,i}^T\Theta_{kl}^T\boldsymbol{A}_l\Theta_{kl}\boldsymbol{z}_{k,i} - \boldsymbol{b}_{kl,i}^T\Theta_{kl}\boldsymbol{z}_{k,i} + \boldsymbol{c}_{kl,i}, \tag{3.9}$$

where

$$\boldsymbol{A}_l = 0.5[\boldsymbol{I}_{D_{xl}} - (1/(D_{xl}+1))\boldsymbol{1}_{D_{xl}}\boldsymbol{1}_{D_{xl}}^T] \tag{3.10}$$

is a constant Hessian matrix, whose entries depend only on the dimension of the observation space. The other intermediate parameters $\boldsymbol{b}_{kl,i}$ and $\boldsymbol{c}_{kl,i}$ are given as follows:

$$\boldsymbol{b}_{kl,i} = \boldsymbol{A}_l\Phi_{kl,i} - \mathcal{S}(\Phi_{kl,i}), \tag{3.11}$$

$$\boldsymbol{c}_{kl,i} = \frac{1}{2}\Phi_{kl,i}^T \boldsymbol{A}_l \Phi_{kl,i} - \mathcal{S}(\Phi_{kl,i})^T \Phi_{kl,i} + \text{lse}(\Phi_{kl,i}), \tag{3.12}$$

where $\Phi_{kl,i}$ is the Taylor series expansion point, which is optimized as a free variational parameter. Note that intermediate parameters are a deterministic function of $\Phi_{kl,i}$. Plugging the approximation in Eqn. 3.9 to Eqn. 3.5 results in a convex lower bound on ELBO, denoted as $\mathcal{L}'$, which is $\leq \mathcal{L}$ and tight at $\Phi_{kl,i}$. Using $\mathcal{L}'$ resolves the intractable integration in Eqn. 3.5, resulting in closed-form posterior parameter estimates, as described in the next section.

Posterior Distributions - E-step

The E-step in the variational EM algorithm computes approximate posterior distributions of the latent variables, which are subsequently used to compute the expectations in Eqn. 3.5. Particularly, there are two parameters to be estimated for each latent variable $\boldsymbol{z}_{k,i}$, which are the mean vector $\boldsymbol{m}_{k,i}$ and the covariance matrix $\boldsymbol{S}_{k,i}$. It is straightforward to maximize over these parameters by using the completing-the-square approach [57] (See Appendix 3.5.1). The terms that quadratically depend on $\boldsymbol{z}_{k,i}$ in the joint log-likelihood yield the posterior covariance update:

$$\boldsymbol{S}_{k,i} = \left[\Sigma_k^{-1} + \sum_{l=1}^{L} N_{kl,i} \Theta_{kl}^T \boldsymbol{A}_l \Theta_{kl}\right]^{-1}, \tag{3.13}$$

where $N_{kl,i}$ is the total number of counts of the ith data sample. Similarly, the terms that linearly depend on $\boldsymbol{z}_{k,i}$ yield the posterior mean update:

$$\boldsymbol{m}_{k,i} = \boldsymbol{S}_{k,i}\left[\Sigma_k^{-1}\boldsymbol{\mu}_k + \sum_{l=1}^{L} (\boldsymbol{x}_{kl,i} + N_{kl,i}\boldsymbol{b}_{kl,i})\Theta_{kl}\right]. \tag{3.14}$$

Lastly, we update the Taylor series expansion point as:

$$\Phi_{kl,i} = \Theta_{kl}\boldsymbol{m}_{k,i}. \tag{3.15}$$

Note that the update of $\Phi_{kl,i}$ depends on the posterior mean. Hence, the algorithm repeats the updates in Eqn. 3.13, Eqn. 3.14, and Eqn. 3.15, respectively, until convergence of the expansion point $\Phi_{kl,i}$.

Point Estimates - M-step

The M-step in the variational EM algorithm maximizes the ELBO with respect to the model parameters. Using the posterior distributions computed in the E-step, we compute the lower bound $\mathcal{L}^{'}$ by taking the expectations with respect to the posterior distributions. Afterward, taking the derivatives with respect to the model parameters yields closed-form update equations for the model parameters. Specifically, the updates for each Θ_{kl} are given as follows:

$$\Theta_{kl} = \Big[\sum_{i=1}^{I_k} \boldsymbol{A}_l^{-1}(\boldsymbol{x}_{kl,i} + N_{kl,i}\boldsymbol{b}_{kl,i})\boldsymbol{m}_{k,i}^T\Big] \Big[\sum_{i=1}^{I} N_{kl,i}(\boldsymbol{m}_{k,i}\boldsymbol{m}_{k,i}^T + \boldsymbol{S}_{k,i})\Big]^{-1}, \tag{3.16}$$

where $\boldsymbol{A}_l^{-1} = 2[\boldsymbol{I}_{D_{xl}} + ((D_{xl}+1)/(D_{xl}+2))\mathbf{1}_{D_{xl}}\mathbf{1}_{D_{xl}}^T]$ using Matrix inversion lemma. The update equations for the mean parameter and covariance of the prior distribution of $\boldsymbol{z}_{k,i}$ then follow as:

$$\boldsymbol{\mu}_k = \frac{1}{I_k}\sum_{i=1}^{I} \boldsymbol{m}_{k,i}, \tag{3.17}$$

$$\Sigma_k = \frac{1}{I}\sum_{i=1}^{I}(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)^T + \boldsymbol{S}_{k,i}, \tag{3.18}$$

respectively. Derivations are given in Appendix 3.6 and the variational EM algorithm is summarized in Algorithm 1.

Limitations

The model parameters of the proposed model are unidentifiable. Due to the Gaussian prior on the latent variables, arbitrary rotation on Θ_{kl} results in the

same likelihood [63]. This makes the direct interpretation of the inferred latent variables ambiguous. This problem can be addressed by enforcing Θ_{kl} to be lower triangular and the main diagonal to be concurrently constrained to be positive [115]. Alternatively, the parameter matrix can be forced to be orthonormal and the columns are ordered by decreasing the variance of the associated latent factors, as in PCA. For more interpretability, sparsity-promoting priors such as ARD [116] and spike-and-slab [117] can be considered, or varimax method [118] can be used to determine the proper rotation matrix. However, identifiability does not affect the predictive performance nor the predictor of covariance, which is the main focus of this paper, hence we leave these constraints in future work.

The objective function optimized by the proposed variational EM algorithm is invariant to rotations, and consequently, the final parameter estimates depend on the initialization. In other words, each initialization of the EM algorithm may result in different parameter solutions, which correspond to different local minima. However, the EM algorithm converges to a stationary point regardless of the initialization since it is guaranteed to monotonically increase $\mathcal{L}'$. One can always improve on the estimator by restarting the EM algorithm multiple times and choosing the maximal converged value. However, we didn't observe significant improvement in covariance prediction accuracy using restarts.

Algorithm 1 Proposed Variational EM algorithm

Input: $\{D_k\}_{k=1:K}$
Initialize $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \{\boldsymbol{\Theta}_{kl}, \{\boldsymbol{\Phi}_{kl,i}\}_{i=1:I_k}\}_{l=1:L}\}_{k=1:K}$
while not $\mathcal{L}^{'}$ converged **do**
  for $k = 1$ to K **do**
    for $i = 1$ **to** I_k **do**
      Infer posterior covariance $\boldsymbol{S}_{k,i}$ by Eqn. 3.13
      Infer posterior mean $\boldsymbol{m}_{k,i}$ by Eqn. 3.14
      for $l = 1$ **to** L **do**
        Update variational parameter $\boldsymbol{\Phi}_{kl,i}$ by Eqn. 3.15
      end for
    end for
    for $l = 1$ **to** L **do**
      Estimate Θ_{kl} by Eqn. 3.16
    end for
    Estimate $\boldsymbol{\mu}_k$ by Eqn. 3.17
    Estimate $\boldsymbol{\Sigma}_k$ by Eqn. 3.18
  end for
  Compute $\mathcal{L}^{'}$ by Eqn. 3.5
end while
Output: $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \{\boldsymbol{\Theta}_{kl}\}_{l=1:L}\}_{k=1:K}$

3.2.4 Model-predicted Density

By virtue of our quadratic surrogate model for the posterior, we can derive an expression for the posterior covariance matrix from the inferred model. The proposed model enables this derivation of the predicted covariance matrix in two ways. First, are a set of Gaussian latent variables, common for all species, which model the observation covariance matrix with a low-rank decomposition. Second, is the quadratic lower bound on the multinomial likelihood, which results in a multivariate Gaussian form for the likelihood of the observations marginalized over the latent variables. Particularly, we define a transformed version of the sample $\boldsymbol{x}_{kl,i}$ as $\tilde{\boldsymbol{x}}_{kl,i}$ with the following function:

$$\tilde{\boldsymbol{x}}_{kl,i} = \boldsymbol{A}_l^{-1}(\boldsymbol{b}_{kl,i} + \boldsymbol{x}_{kl,i}), \tag{3.19}$$

where $\boldsymbol{A}_l$ is the matrix defined in Eqn. 3.10. Then, it is straightforward to show that the likelihood of the transformed data $\tilde{\boldsymbol{x}}_{kl,i}$ is given as follows (See Appendix

3.6):

$$
\begin{aligned}
& p(\tilde{\boldsymbol{x}}_{kl,i}; \Theta_{kl}, \boldsymbol{\mu}_k, \Sigma_k) \\
&= \int \mathcal{N}(\tilde{\boldsymbol{x}}_{kl,i}; \Theta_{kl} \boldsymbol{z}_{k,i}, \boldsymbol{A}_l^{-1}) \mathcal{N}(\boldsymbol{z}_{k,i}; \boldsymbol{\mu}_k, \Sigma_k) d\boldsymbol{z}_{k,i} \\
&= \mathcal{N}(\tilde{\boldsymbol{x}}_{kl,i}; \Theta_{kl} \boldsymbol{\mu}_k, \boldsymbol{A}_l^{-1} + \Theta_{kl} \Sigma_k \Theta_{kl}^T),
\end{aligned} \tag{3.20}
$$

where the covariance matrix $\boldsymbol{C}_{kl,\text{intra}} = \boldsymbol{A}_l^{-1} + \Theta_{kl} \Sigma_k \Theta_{kl}^T$ and the mean vector $\boldsymbol{\phi}_{kl} = \Theta_{kl} \boldsymbol{\mu_k}$ are of interest to us, in which $\boldsymbol{C}_{kl,\text{intra}}$ captures intra-species correlations of species l in condition k. To obtain inter-species correlations, define $\tilde{\boldsymbol{A}}^{-1} = \text{diag}(\boldsymbol{A}_1^{-1}, \ldots, \boldsymbol{A}_L^{-1})$ and $\tilde{\Theta}_k = [\Theta_{k1}, \ldots, \Omega_{kL}]$, then $\boldsymbol{C}_{k,\text{inter}} = \tilde{\boldsymbol{A}}^{-1} + \tilde{\Theta}_k \Sigma_k \tilde{\Theta}_k^T$ gives a covariance matrix for both inter-species and intra-species. To convert any covariance matrix to a proper correlation matrix, which is useful for visualization and analysis, one can use the transformation $\text{Corr} = \text{diag}(\boldsymbol{C})^{-1/2} \boldsymbol{C} \text{diag}(\boldsymbol{C})^{-1/2}$.

3.2.5 Computational Complexity

The computational complexity of the variational EM algorithm determines the algorithm's scalability to large datasets. For notational simplicity, we assume that there is only one discrete condition, hence we use I instead of I_k. In the E-step, Eqn. 3.13 computes posterior covariance, which requires multiplication of a $d_z \times d_l$ matrix with its transpose resulting $O(d_z^2 d_l)$ complexity. This process is repeated for each species resulting in $O(L d_z^2 d_l)$. Inverting the matrix for each sample costs $O(I d_z^3)$. Hence, the overall asymptotic complexity for the posterior covariance computation is $O(I(d_z^3 + L d_z^2 d_l))$. The posterior mean computation in Eqn. 3.14 involves matrix-vector multiplications that require $O(L d_z d_l)$, and $O(d_z^2)$ due to covariance posterior covariance multiplication. Hence, the total cost per sample is $O(L d_z d_l + d_z^2)$ and the overall cost is $O(I(L d_z d_l + d_z^2))$. Consequently, the complexity of the E-step is $O(I(L d_z d_l + d_z^2 + d_z^3 + L d_z^2 d_l))$. Removing non-

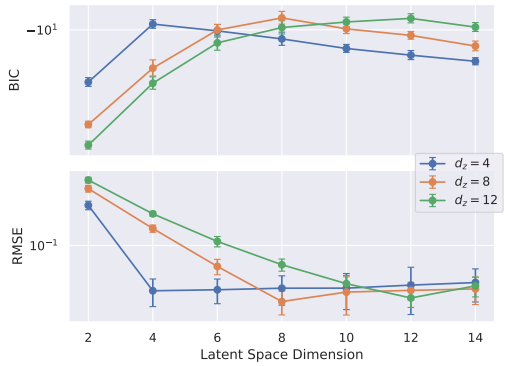
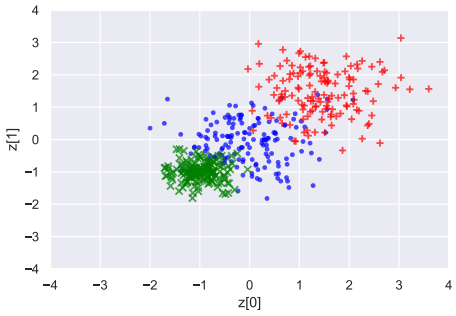


Figure 3.2: BIC approximation to the evidence, and RMSE of the predicted covariance matrix with respect to the latent space dimension d_z. True dimensions are 4, 8, and 12. Blue, orange, and green curves show RMSE and the BIC penalized log likelihood (BIC), respectively. Note that the BIC exhibits a clear maximum over latent space dimension d_z. BIC values are scaled by factor 10^{-3}.

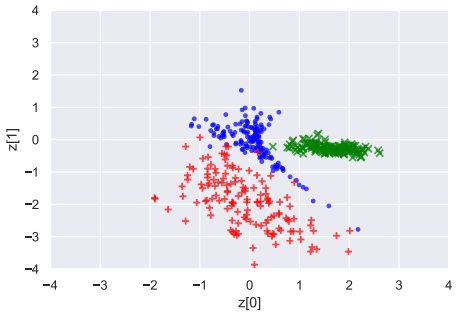
dominant terms results in $O(I(d_z^3 + Ld_z^2 d_l))$. One can see that this scales linearly in terms of L, d_l, and I. On the other hand, the dominant computation in the M-step is for Θ_{kl}. Eqn. 3.16 comprises two terms. The first term requires $O(Id_l d_z)$ due to I times vector-vector outer products. The second term requires $O(Id_z^2 + d_z^3)$ due to vector-vector outer products and subsequently matrix inversion. Multiplying these terms costs $O(d_l d_z^2)$, hence resulting total complexity of $O(L(Id_l d_z + Id_z^2 + d_z^3 + d_l d_z^2))$ for all $l = 1 : L$. It is also clear that this computation scales linearly in terms of L, d_l, and I. Modeling the conditions independently also induces linear complexity in terms of K. In summary, both E and M steps scale linearly in terms of K, L, d_l, and I, which suggests that the proposed optimization algorithm is scalable for large datasets as long as the latent space dimension d_z is relatively small.

3.3 Experiments

In this section, we perform numerical experiments to illustrate the proposed model. We start with simulation studies, then conclude with experiments on a



(a) True embeddings



(b) Proposed model embeddings

Figure 3.3: 2D Latent space visualization of 100D count vectors

bacterial microbiome dataset.

3.3.1 Simulations

We generate synthetic datasets i) to explain the model selection strategy, ii) to demonstrate the accuracy of the latent embeddings, and iii) to show the ability to capture the covariance structure from observed data.

Model Selection

The proposed algorithm estimates the covariance matrix with a low-rank decomposition. The rank of the matrix is equal to the number of components d_z in the latent space. which is a model hyper-parameter to be determined. We use the Bayesian Information Criterion (BIC) to estimate this parameter using only the training dataset. The BIC arises from the Laplace approximation to the model

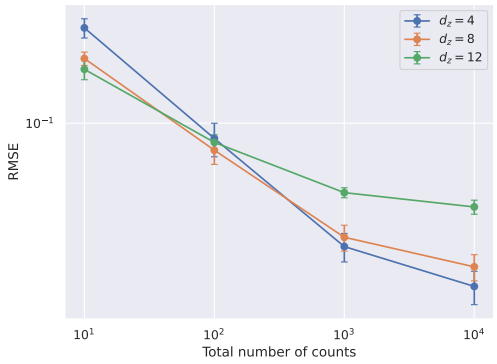


Figure 3.4: RSME of the covariance estimation with respect to the average total number of counts observed in the metatranscriptomic data for different latent space dimensions d_z. As the counts increase the errors decrease until a saturation limit. The lower the dimension of the latent space, the more sensitivity to the total number of counts.

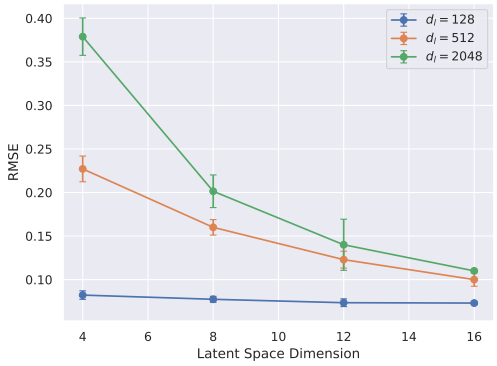
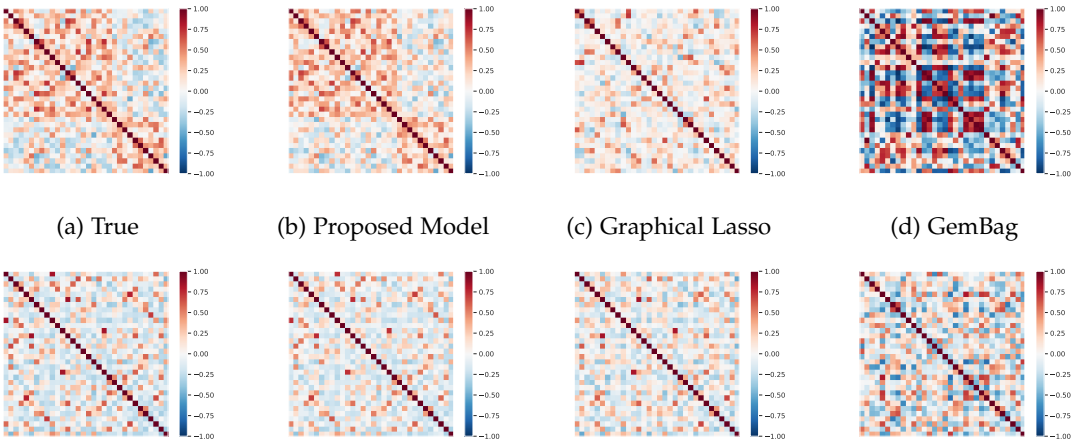


Figure 3.5: RMSE of the predicted covariance matrix with respect to the latent space dimension for three different observation space dimensions d_l.

posterior $p(M|D_k)$ [119], where M is the complete model including the latent dimension d_z. This results in a Bayesian estimate of d_z: $d_z = \text{argmax}_{d_z}\text{BIC}$, where $\text{BIC} = \log p(D_k) - 0.5 \times \log I_k \times \text{dof}$, which is a function of the total number of unknown parameters that penalizes the log-likelihood with a model complexity penalty term. In the proposed model, we use ELBO lower bound to the likelihood by following [120]. The unknown parameters of the model are $\{\Theta_{kl}\}_{l=1}^{L}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$. Hence, the total number of parameters is $\text{dof} = K \times (d_l + K)$. To illustrate the BIC model selection for the proposed model, we simulate three datasets with true latent space dimensions 4, 8, and 12, respectively, and then train multiple models



(a) True (b) Proposed Model (c) Graphical Lasso (d) GemBag

(e) Emprical (f) Ledoit-Wolf (g) FA (h) jSDM

Figure 3.6: Estimated normalized covariance matrices produced by the considered algorithms for a two-species community with simulated transcript data. For more details on the model see Section 3.3.1. The proposed model provides a much more accurate estimated covariance than the other methods.

while varying the dimensions d_z over $\{2, 3, \ldots, 12\}$ as the search range. We repeat the experiment 50 times to report the performance. The panel on top of Fig. 3.2 shows the average BIC values obtained after convergence of the variational EM algorithm. We see that maximum BIC values are obtained in the vicinities of the true ranks for all the datasets. On the other hand, the panel on the bottom shows the RMSE values of the estimated covariance matrices. One can see that the lowest errors are achieved at the ground truth d_z values, which validates the model selection method.

Embedding Characteristics

We generate a synthetic dataset with a 2-dimensional latent space having 3 different classes, i.e., experimental conditions, according to the model specification in Section 3.2.2. The latent variables are sampled for each class from different Gaussian distributions. The associated means are predefined as $[0, 0]$, $[1.5, 1.5]$, $[-1, -1]$, and the variances of the isotropic covariances are selected as $0.5, 0.5, 0.1$, respec-

tively. Three class conditional densities are generated with different affine transformation parameters. The observation space is 25-dimensional. The observations are sampled from the conditional multinomial distributions with soft-max link function as in Eqn. 3.2. We generate 200 observations for each class with a fixed total number of counts, which is 100, per observation, then stack all the observations. Fig. 3.3.a shows the true embeddings of the resulting dataset. We trained the proposed algorithm with the true latent space dimension. Fig. 3.3.b shows the embeddings of the model, which are obtained through the posterior distributions. Due to the non-identifiability of the model, the latent variables can only be recovered up to a rotation. The distorted shape of the latent clusters in Fig. 3.3.b is due to the use of the soft-max link function. If there is a large component in the affine transformed latent vector, the other components are washed out, hence such points would map to very close points in the observation space. Notwithstanding the differences between Fig. 3.3.a and Fig. 3.3.b, the model preserves the clustering structure accurately.

Influence of the Total Counts and Dimensions on Performance

The number of counts of the observed vector $\boldsymbol{x}_{kl,i}$ is an observation-specific parameter, which affects the accuracy of the proposed algorithm. Figure 3.4 shows the effect of the number of counts $N_{kl,i}$ on the RMSE values of the covariance estimator under three different latent dimension settings. We sample the total counts of a simulated vector from the Poisson distribution with a fixed mean. We also fix the observation dimension to 128. RMSE is reported based on averaging 50 experiments. Figure 3.4 shows that increasing the mean number of counts improves performance. In particular, we see that the total counts $N_{kl,i}$ and the

mean error are inversely proportional. This is expected since the number of counts directly affects the posterior uncertainty (Eqn. 3.13) and mean (Eqn. 3.14). The contribution to the ELBO of the observations increases as the total count increases. Furthermore, for the low number of counts, the covariance matrix becomes harder to predict due to higher vulnerability to over-fitting. Hence, the lower the dimension of the latent space, the more sensitivity to the total number of counts. On the other hand, Fig. 3.5 demonstrates the opposite trend when the dimension of the observation space dimension is increased. Here the total mean counts are fixed at 1000. In higher dimensional datasets, the model struggles to estimate the covariance structure when the rank is low. However, this phenomenon diminishes when we observe more counts as can be seen in Fig. 3.4.

Baseline Algorithms

Here we present the performance comparisons of the proposed method relative to several baseline methods for estimating the underlying covariance and inverse covariance matrices. i) Empirical covariance, which is computed as the sample covariance. ii) The Ledoit-Wolf estimator [121], which uses shrinkage regularization to perform MAP estimation for the covariance matrix by assigning an inverse Wishart prior to the covariance matrix. iii) Gaussian Copula GraphicalLasso [66], which penalizes the precision matrix with L1-norm constraints after transforming the data by using Gaussian copulas. Regularization forces the entries of the precision matrix to be sparse. iv) Factor Analysis [63] uses another form of regularization of the covariance matrix by imposing a low-rank structure. v) GemBag [78], assumes a common sparsity structure among the environmental conditions by modeling the edges of the environment-specific precision matri-

ces using hierarchical priors. Each of these aforementioned baseline methods is expected to perform best when the data, or its transformed version, is normally distributed. vi) jSDM [80], which is another latent variable model that uses the log link function for the count observations. Environmental conditions and latent variables are graphically joined at the mean vectors, i.e., logits. For the Empirical Covariance, Ledoit-Wolf, GraphicalLasso, FA, and GemBag, we first normalize the data by subtracting the mean and dividing by the variance, before running these methods. On the other hand, the non-Gaussian counting nature of the data is explicitly modeled in our proposed model and jSDM, thus running on the raw observations. For model selection in FA, jSDM, and the proposed model, we use the exact rank of the simulated dataset. The regularization coefficient of the Gaussian Copula GraphicalLasso algorithm is estimated by using 5-fold cross-validation. For the Ledoit-Wolf algorithm, we used the expression for the shrinkage coefficient given in [121]. See Section III in the supplementary for further implementation details.

Simulating Model Communities

Next, we generate a synthetic dataset that contains the transcript abundance data (an estimate of gene expression) of two different species existing in the same community, hence $L = 2$. The latent variables $\boldsymbol{z}_i$ with dimension $d_z = 3$ are generated for each measurement site by sampling from $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}_{d_z}, \boldsymbol{I}_{d_z})$, where i indexes the replicate for $i = 1, \ldots, I$. These latent variables have elements that correspond to the hidden factors generating the data, such as environmental variables, mediator species effects, and direct associations. We transform the latent variables to the probabilities in the observation space, whose dimensions

(abundance of transcripts) are chosen as $d_1 = 20, d_2 = 10$ by using affine and subsequently soft-max transformations as described in Section 3.2.2. The parameters $\Theta_l \in \mathbb{R}^{d_l \times d_z}$, are chosen randomly by sampling from a zero mean multivariate normal distribution. Then, we sample the observed data $\boldsymbol{x}_{l,i}$ from the multinomial distribution. The total counts $N_{l,i}$ of a sample is chosen randomly by sampling from a Poisson distribution with rate parameter 1000. We simulate a total of $I = 100$ replicates for each environmental condition where the total number of conditions $K = 2$. The true covariance matrix is then given as $\tilde{\Theta}_k \tilde{\Theta}_k^T$, where $\tilde{\Theta}_k = [\Theta_{k,1}, \Theta_{k,2}]$.

Correlation Results

Fig. 3.6 shows the estimated covariance matrices of the baseline algorithms, the proposed algorithm, alongside the ground truth matrix, when the simulated datasets are realized following Section 3.3.1. The proposed model can recover the covariance structure accurately. The relatively poorer accuracy of the other methods can be attributed to several factors. First, these models do not exploit the counting nature of the data. The second reason is that the covariance matrix is simulated with a low-rank structure, which is not taken into account by the Gaussian Copula Graphical-Lasso, Ledoit-Wolf, or standard sample covariance estimation methods. Third is the common structure assumption of GemBag and jSDM among the covariance/precision matrices for each environment. As the data were simulated from the proposed model, the proposed algorithm naturally performs better. Section IV in the supplementary discusses more on model mismatch concept with additional simulations. Note also that, for multiple species, the proposed model can discover both inter-species and intra-species

Table 3.1 Mean and standard deviation of RMSE between the estimated covariance/precision matrices and ground truths over 50 different realizations of the simulated abundance dataset.

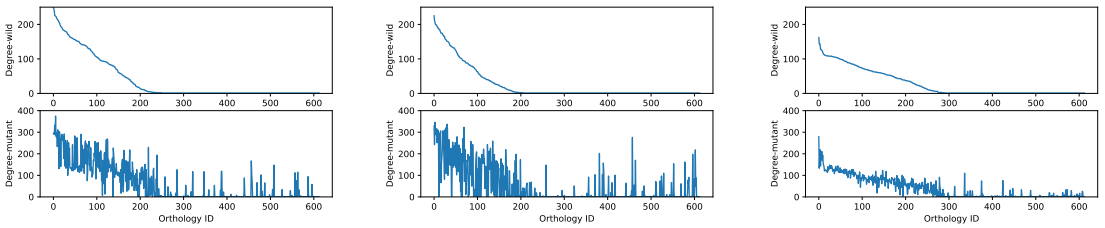
Algorithm	Covariance	Precision
Empirical	$.366 \pm .012$	$.291 \pm .016$
Ledoit-Wolf	$.340 \pm .013$	$.142 \pm .002$
GLasso	$.343 \pm .015$	$.090 \pm .002$
GemBag	$.278 \pm .009$	$.102 \pm .002$
FA	$.317 \pm .036$	$.159 \pm .007$
jSDM	$.310 \pm .025$	$.071 \pm .004$
Proposed	$.176 \pm .016$	$.023 \pm .001$

correlations. Table 3.1 shows the resulting RMSE values between the estimated and the ground truth covariance matrices for the aforementioned simulation setting. The proposed model achieves lower error overall. This is expected since the model uses an ELBO approximation to the true marginal likelihood function.

3.3.2 Bacterial Community Experiment

In this section, we demonstrate a real-world use-case of the proposed model: transcript analysis of a bacterial model community called THOR [122].

Microbial model communities are useful to understand principles that govern community behaviors [123, 124, 125, 126]. The Hitchhikers Of the Rhizosphere (THOR) is a model community consisting of three microbial species, *Bacillus cereus*, *Flavobacterium johnsoniae*, and *P. koreensis* that co-isolate from field-grown soybean roots. The organisms in THOR represent three dominant rhizosphere taxa (at the phylum level), and are common in soil and the mammalian gut. *B. cereus* is a Firmicute that carries *F. johnsoniae*, a member of the Bacteriodetes, and *P. koreensis*, a member of the Proteobacteria, as hitchhikers [127]. Due to their abundance in several environments, their may demonstrated interactions in the lab and field, and their genetic tractability, these species make a useful model community with relevance to the natural world. The model community provides a simple system in which to study and model community-level interactions, which are poorly



(a) B. cereus (b) F. johnsoniae (c) P. koreensis

Figure 3.7: Effect of koreenciene removal on the centrality (vertex degree) of vertices in the transcriptional orthology correlation networks inferred from our model for the experimental THOR dataset. For each species, the ortholog IDs are sorted in decreasing order of the wildtype vertex degree. The upper row shows plots of the degree of each vertex (transcriptional orthology ID), in descending order of magnitude, for the wildtype condition. The bottom row shows corresponding plots of the vertex degree when the koreenceine pathway is removed (mutant condition), under the same ordering of vertices as in the top row. *P. koreensis* preserves its network connectivity better than the other two species. The network connectivity of *F. johnsoniae* is the most affected by koreenciene removal.

understood. Developing governing principles of community behavior may lead to strategies to manipulate microbiomes for human or environmental health.

The dataset is collected under two conditions associated with the treatments applied to *P. koreensis*. In the first condition, the THOR community contains the wild type *P. koreensis* strain and in the second condition the wildtype is replaced with a mutant of *P. koreensis* that does not produce koreenceine antibiotics. Production of koreenceines is an important factor in community interactions because they inhibit the growth of *F. johnsoniae* [128] and *B. cereus* protects *F. johnsoniae* by modulating koreenceine levels. By using our proposed model, in particular the associated estimated joint probability density of the data, we will be able to reveal the effects of the treatment. Since the joint probability density model is parameterized by the mean and covariance of a multivariate Gaussian latent variable (See Section 3.2.4), the mean and covariance parameters play the principal role in our metatranscriptomic analysis. For brevity, we focus our discussion on the inferred covariance parameters here (See Supplementary for discussion of the

mean parameters inferred by the model).

The microbial community dataset consists of a total of 17244 gene transcripts associated with three species. There were respectively 38 and 36 replicates for the community with wildtype and mutant strains of *P. koreensis*. 343 transcripts were removed from the analysis as they had zero counts over all experimental replicates. After removing these transcripts, *B. cereus, F. johnsoniae,* and *P. koreensis* express 5903, 5146, 5852 transcripts, respectively. We reduced the dimension of the feature space using orthological groupings of gene transcripts into metabolic pathways[1]. Specifically, after pathway mapping, each feature corresponds to a transcriptional orthology ID, and the associated data is the summation of the counts of the transcripts tagged with that ID. We aggregated all the transcripts that were not mapped to any Kegg ortholog into a single non-assigned orthology ID, denoted KXXXXX, and we only considered those ortholog IDs that are present in all 3 species. This filtering resulted in a set of 613 ortholog IDs, which corresponds to the dimension of the feature space used in our model.

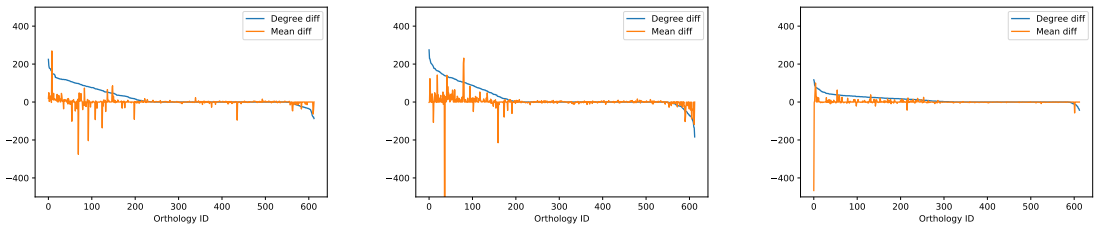
The rank of the proposed model was determined by successively fitting the model to latent spaces of dimensions ranging between 5 and 50 with increments of 5. Then, the optimal model rank was determined as the latent dimension that yields the highest value of the BIC as described in Section 3.3.1. The optimal model rank was found to be 40. The parameters (mean and covariance) of the models were subsequently refitted with the optimal dimension. The probability distribution of the data is computed under the wildtype and mutant conditions, whose explicit form is given in Eqn. 3.20 as a marginalization over the latent

[1]The transcriptional orthology mappings of the THOR gene transcripts to metabolic pathways were obtained using Kegg `https://www.genome.jp/kegg/`. See supplementary for an example.

variables.

Network centrality changes: We evaluate the effect of the removal of koreenciene (mutant) on the centrality of the inferred 613×613 correlation network of metabolic pathways. Here the centrality of a vertex of the network is measured by vertex degree, i.e., the number of edges connecting the vertex. To ensure that the networks contain only the most biologically significant edges in the networks, we applied a very high correlation threshold (0.95) to the respective inferred wild-type and mutant correlation matrices produced by fitting our proposed graphical model to the data. Using such a high threshold is in line with established RNA-Seq network inference practices [129]. Figure 3.7 illustrates the effect of the removal of koreenceine on the degrees of the nodes (transcriptional orthology IDs) in these networks. Comparison of the upper panels with the lower panels of the figure indicates that the vertex degree distribution of *F. johnsoniae* is most affected, followed by *B. cereus*, with *P. koreensis* the least affected. This relative ordering of sensitivity of the three species to koreenceine removal shown for vertex degree in Fig. 3.7 mirrors the relative ordering of sensitivity shown for the mean changes (See Fig. 2 and associated discussion in the Supplementary).

Fig. 3.7 illustrates the relative effect of koreenceine removal on increases vs decreases in vertex degree of the transcriptional orthology correlation network for each species. In the figure, the transcriptional orthology IDs are sorted according to the difference between mutant vs wildtype vertex degree. The blue curve shows the resultant vertex degree difference and the orange curve shows the vertex mean difference. Observe that the order of decreasing differences of vertex degree does not correspond to the order of decreasing differences in vertex mean.



(a) B. cereus (b) F. johnsoniae (c) P. koreensis

Figure 3.8: Effect of koreenceine removal on vertex centrality and vertex mean counts for the transcriptional orthology correlation network. For each species, the ortholog IDs are sorted in decreasing order of the vertex degree difference between mutant and wild type. It is notable that, with few exceptions, all orthology IDs with significant changes in vertex mean also have changes in vertex degree, but not conversely. Furthermore, the asymmetry of the blue curve suggests that the removal of koreenceine is associated with an increase in network connectivity (many more vertices whose degrees increase than decrease), especially in *F. johnsoniae*.

However, a change in the vertex mean almost always accompanies a change in vertex degree, although the converse is not true. Also note from the asymmetry of the blue curves in Fig. 3.7 that the mutant's networks have many more vertices that increase than decrease in vertex degree as compared to the wild type. Thus koreenceine removal seems to increase network centrality of a large number of transcriptional orthologs, especially for *F. johnsoniae*. We point out that the large spikes that appear in the orange curves (vertex mean difference) for *F. johnsoniae* and *P. koreensis*, correspond to the ID KXXXXX, which are genes that were not mapped to any Kegg transcriptional ortholog. Further discussion can be found in the supplementary.

In summary, the proposed model can provide two important data analysis components for microbiome model community analysis. First, we can assess transcriptional orthology composition changes under the treatment by observing the means of the marginal distributions provided by the proposed model. Second, we can assess the second-order interaction changes by using the correlation networks that are obtained from the covariance matrices of the marginal distributions.

These two components along with the abundance ratio analysis in [122] provide a complementary analysis of microbial model communities, which can further be interpreted by microbiologists.

3.4 Conclusion

A hierarchical Bayesian latent variable model was proposed for the joint analysis of multiple discrete datasets. We explained the associations between the features of the datasets with a common lower dimensional latent space, represented by a set of independent identically distributed Gaussian random variables. To overcome the lack of conjugacy between the multinomial observation distribution and the Gaussian latent space distribution, we developed a variational EM algorithm based on quadratic bound approximations for estimating the parameters in the model. The computation of the algorithm scales linearly with the number of features, samples, and datasets. Simulation studies show that the proposed model can recover low-rank covariance structures accurately. Furthermore, our real-world microbiome experiment demonstrates the potential real-world utility of the model for the exploration of correlation and associated networks for dichotomous microbiome data.

There are several promising directions for future work. One possible area is to generalize the model to capture covariance structures of absence-presence datasets by modeling the binary observations using Bernoulli distributions. Another generalization can be achieved by the incorporation of covariates such as temperature, ph, and physical/chemical perturbations, that may change the composition of the species. The mean of the latent variables can be made a function of the covariates to accomplish that. One another possible area is to incorporate system dynamics

into the latent space so as to explicitly capture temporal correlations. In particular, there is increasing interest in collecting longitudinal microbiome data for studying adaptation, resilience, and dynamics over time. The incorporation of a state-space dynamical model into our framework can reveal the temporal evolution of the interactions between the genomes. Another future direction is to improve the parsimony of the model by incorporating sparsity into the latent representation by using sparsity-inducing priors for the covariance or inverse covariance (precision) matrices.

3.5 Appendix

3.5.1 Estimation of Posterior Parameters

Log-likelihood of Multivariate Normal distribution $\log \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as:

$$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const$$

in which the second order term in $\boldsymbol{x}$ corresponds to the inverse of covariance matrix $\boldsymbol{\Sigma}$, and the linear term corresponds to the mean when multiplied with $\boldsymbol{\Sigma}$. Inferring the mean and covariance from linear and quadratic terms is called completing the square approach. We make use of this method to infer the posterior distributions of $z_{k,i}$, which is denoted as $q(z_{k,i}; \boldsymbol{m}_{k,i}, \boldsymbol{S}_{k,i})$. Given the joint likelihood in Eqn. 3.6 and quadratic approximation in Eqn. 3.9, one can collect the quadratic terms in $z_{k,i}$ as follows:

$$-\frac{1}{2}z_{k,i}^T\boldsymbol{\Sigma}_k^{-1}z_{k,i} - \sum_{l=1}^{L}\frac{N_{kl,i}}{2}z_{k,i}^T\Theta_{kl}^T\boldsymbol{A}_l\Theta_{kl}z_{k,i},$$

which follows Eqn. 3.13 for posterior covariance $\boldsymbol{S}_{k,i}$ estimate. Similarly, the linear terms are collected as:

$$\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\boldsymbol{z}_{k,i} + \sum_{l=1}^{L} \boldsymbol{x}_{kl,i}\boldsymbol{\Theta}_{kl}\boldsymbol{z}_{k,i} + N_{kl,i}\boldsymbol{b}_{kl,i}^T\boldsymbol{\Theta}_{kl}\boldsymbol{z}_{k,i}.$$

Collecting the terms and multiplying with the posterior covariance estimate yields posterior mean $\boldsymbol{m}_{k,i}$ estimate as given in Eqn. 3.14.

3.5.2 A note on Quadratic Surrogate Optimization Transfer

Due to the non-conjugacy between multinomial and multivariate normal distributions, computing the posterior distributions of the latent variables is intractable, hence we can not obtain closed form expressions for the expectations of the joint likelihood required for the M-step. We adopt an alternative variational inference approach, called quadratic surrogate optimization transfer, where the problematic terms of the joint log-likelihood are replaced with simpler quadratic surrogates obtained by truncated Taylor series expansion. These quadratic functions have tunable free variational parameters and expansion points that control the tightness of the approximation, which are optimized concurrently. This differs from the mean-field approach of variation Bayes inference, which is a global approximation that uses a factorized approximation to the multivariate posterior distribution in order to make the computation of statistical expectation tractable. Quadratic surrogate optimization transfer is on the other hand performed locally for the problematic terms of the joint likelihood, i.e., the approximated quadratic function is created and optimized at each iteration of the EM algorithm. In the literature, this approach has been used for logistic regression [86], multi-task learning [88], discrete factor analysis [90], and correlated topic models [62].

3.5.3 Upper bound to *LogSumExp* Function

For notational simplicity, let $\boldsymbol{\eta} = \Theta_{kl} \boldsymbol{z}_{k,i}$ and drop k, l, and i indexes. Second order Taylor series expansion at arbitrary point $\boldsymbol{\psi}$ yields:

$$\text{lse}(\boldsymbol{\eta}) \leq \text{lse}(\boldsymbol{\psi}) + \mathcal{S}(\boldsymbol{\psi})^T(\boldsymbol{\eta} - \boldsymbol{\psi}) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\psi})^T \boldsymbol{A}(\boldsymbol{\eta} - \boldsymbol{\psi}),$$

where we replace the original Hessian matrix with constant $\boldsymbol{A}$ in Eqn. 3.10 to obtain the upper bound based on [86]. Reorganizing the terms corresponds to the following quadratic function:

$$\frac{1}{2}\boldsymbol{\eta}^T \boldsymbol{A}\boldsymbol{\eta} - (\boldsymbol{A}\boldsymbol{\psi} - \mathcal{S}(\boldsymbol{\psi}))^T \boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\psi}^T \boldsymbol{A}\boldsymbol{\psi} - \mathcal{S}(\boldsymbol{\psi})^T \boldsymbol{\psi} + \text{lse}(\boldsymbol{\psi}).$$

Then, we introduce $\boldsymbol{b}$ and $\boldsymbol{c}$ for linear and constant terms, respectively, to simplify the notation in the main text.

3.6 Derivation of M-step Updates

Taking expectation of $\mathcal{L}'$ with respect to the posterior distributions of the latent variables in Eqn. 3.14 and Eqn. 3.13 yields the following expression:

$$\begin{aligned}
&\sum_{i=1}^{I_k} \big[\boldsymbol{x}_{kl,i}^T \boldsymbol{\Theta}_{kl} \boldsymbol{m}_{k,i} - \frac{N_{kl,i}}{2} \boldsymbol{m}_{k,i}^T \boldsymbol{\Theta}_{kl}^T \boldsymbol{A}_l \boldsymbol{\Theta}_{kl} \boldsymbol{m}_{k,i} \\
&+ N_{kl,i} \boldsymbol{b}_{kl,i}^T \boldsymbol{\Theta}_{kl} \boldsymbol{m}_{k,i} + \frac{1}{2} \log |\boldsymbol{\Sigma}_k^{-1}| \\
&- \frac{1}{2}(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{m}_i - \boldsymbol{\mu}_k) - \frac{1}{2} Tr(\boldsymbol{\Sigma}_k^{-1} \boldsymbol{S}_{k,i}) \\
&- \frac{1}{2} N_{kl,i} \text{vec}(\boldsymbol{\Theta}_{kl})^T (\boldsymbol{A}_l \otimes \boldsymbol{S}_{k,i}) \text{vec}(\boldsymbol{\Theta}_{kl}) \big] + const,
\end{aligned}$$

where vec denotes the vectorization, Tr denotes the trace operator, and $\otimes$ is the Kronecker product. The derivatives of this expression with respect to $\boldsymbol{\Theta}_{kl}$, $\boldsymbol{\mu}_k$, and

$\boldsymbol{\Sigma}_k^{-1}$ are given respectively as follows:

$$\boldsymbol{\Theta}_{kl} \rightarrow \sum_{i=1}^{I_k} \big[\boldsymbol{x}_{kl,i}\boldsymbol{m}_{k,i}^T + N_{kl,i}\boldsymbol{b}_{kl,i}\boldsymbol{m}_{k,i}^T - N_{kl,i}\boldsymbol{A}_l\boldsymbol{\Theta}_{kl}\boldsymbol{m}_{k,i}\boldsymbol{m}_{k,i}^T - N_{kl,i}\boldsymbol{A}_l\boldsymbol{\Theta}_{kl}\boldsymbol{S}_{k,i}\big],$$

$$\boldsymbol{\mu}_k \rightarrow \sum_{i=1}^{I_k} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k),$$

$$\boldsymbol{\Sigma}_k^{-1} \rightarrow \frac{1}{2}\boldsymbol{\Sigma}_k - \frac{1}{2}\sum_{i=1}^{I_k}[(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)^T(\boldsymbol{m}_i - \boldsymbol{\mu}_k) + \boldsymbol{S}_{k,i}].$$

Equating the derivatives to zero results in closed form update equations in Eqn. 3.16, Eqn. 3.17, and Eqn. 3.18, for $\boldsymbol{\Theta}_{kl}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, respectively.

CHAPTER IV

Hierarchical Bayesian Multitask Logistic Regression Model for Microbiome Profiling

4.1 Introduction

Multitask learning (MTL) is a class of machine learning prediction models where multiple related learning tasks are trained jointly [130] (see [131] for a recent survey). This allows us to combine multiple related tasks (datasets) together to increase the effective sample size, while keeping the interpretability of a single base model. MTL has been an active area of research with applications including: face recognition in computer vision; joint analysis of heterogeneous genomics data; and social media sentiment analysis [132, 133, 134, 135].

In this chapter, we introduce a hierarchical Bayesian multitask logistic regression model tailored for binary predictions on multiple related datasets. While the model is broadly applicable, we focus on the predictions of the disease conditions of patients based on gut microbiome data, specifically.

It has been well established that microbes within the human gut affects human health [136, 137, 138]. Motivated by the success of machine learning models in areas such as computer vision, medical imaging, and protein prediction [139, 140, 141], there has been an increasing interest to employ machine learning models to

perform health prediction based on human gut data [142, 143, 144, 145]. However, application of machine learning methods to microbiome data faces two major challenges. First typical microbiome data lies in high feature dimension, and the number of microbes is significantly more than the number of samples available [142]. In this regime, overparameterized models can suffer from overfitting of the training data [146]. Second, in health applications it is essential that machine learning models be interpretable and quantify uncertainty in their predictions, which is not the case with most deep neural nets, for example [147].

We propose an interpretable predictor model that is based on a hierarchical Bayesian generalized linear model (GLM) [148]. The model introduces a set of binary variables shared across different datasets to represent the most informative features (i.e bacterial species) for predictions. This enables the model to identify the common bacterial species shared across the different studies where each one corresponding to a distinct pathology.

In contrast to optimization based MTL approaches[149, 150, 151, 152], our proposed Bayesian hierarchical modeling provides natural uncertainty quantification through the posterior distribution of the label given the features and a flexible framework to incorporate domain experts' knowledge [41]. The proposed model differs from [153, 154, 149, 150, 152, 155] in how the sparsity pattern is modeled: inspired by [155] we use an overparameterized Bernoulli-Gaussian model instead of regularizations, which has been demonstrated to have better support recovery properties [156]. The Bayes posterior distribution is not in analytical closed form and we propose an approximation to the posterior mode that is based on variational inference [55, 157], which is more salable to the high feature dimen-

sion characteristic of microbiome datasets. . Variational inference is often less computationally difficult than Monte Carlo methods for evaluating the posterior mode.

We illustrate our model capabilities through numerical experiments in simulation and in a real world microbiome dataset that is composed of 21 different studies[158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176] curated by Lawrence Livermore National Laboratory (LLNL) staff members. Each study from the data contains a collection of patients' gut bacteria sequencing data and the health condition of the patients. The sequencing data undergo a clustering process to form Operational Taxonomic Units (OTUs), each representing a group of similar sequences. These OTUs are then taxonomically classified at 7 different levels depending on the resolution of the data. Their relative abundances (counts) are subsequently quantified across samples from different patients. These abundance data are further processed by centered log ratio transform (CLRT) [177] to provide features for the machine learning model. There are total 7 different taxon levels, and we take the union set of all the OTUs for each taxon level to provide a consistent feature space. The taxon levels are: Kingdom (11), Phylum (163), Class (313), Order (914), Family (2544), Genus (7885) and Species (31518), where each number in the parentheses corresponds to the feature dimension (i.e total number of OTUs). The health condition of each patient is one-hot coded into one of the 11 disease: cirrhosis, inflammatory bowel disease, diabetes, diarrhea, cancer, dermatologic, premature born, cardiovascular disease, neurological disease, gastrointestinal infection and autoimmune. This maps the studies of distinct diseases into different tasks.

The proposed model is evaluated in comparison with other benchmark methods including: logistic regression with sparsity penalty [178, 179], MTFL (Multitask Feature Learning) [149] and MSSL (Multitask Sparsity Structure Learning) [151]. Our evaluation with simulated synthetic datasets show that the proposed approach has superior support recovery property when the underlying regression coefficients share a common sparsity structure across different tasks. The proposed model performs less well on the real microbiome data, likely due to heterogeneity of the data (i.e different experimental objectives, laboratory setups, sequencing equipments, patient demographics etc.), Nonetheless, we demonstrate the utility of the method to extract informative taxons while providing well-calibrated predictions with uncertainty quantification.

The chapter is organized as: Section 4.2 introduces the mathematical formulation of the proposed hierarchical Bayesian model, Section 4.3 presents the proposed variational inference algorithm to obtain the approximated posterior distribution, Section 4.4 provides application of the methods to synthetic datasets and the microbiome dataset, and Section 4.5 summarizes the findings from the chapter and discusses future directions.

4.2 Hierarchical Bayesian Multitask Logistic Regression Model

4.2.1 Notations and Terminologies

We refer to each Operational Taxonomic Unit (OTU) as a feature variable, each study is referred to as a task and diseased or healthy state of the individual is referred to as a label.

We use bold upper case letters for matrices, bold lower case letters for vectors and no bold lower case for scalars. We denote the observed OTU count data after

centered log ratio transform as $\boldsymbol{x}_t^i \in \mathbb{R}^d$ together with its label $y_t^i \in \{0, 1\}$, where d corresponds to the number of features (i.e number of OTUs), $t = 1, \cdots, T$ denotes the different tasks (i.e different studies), $i = 1, \cdots, n_t$ denotes the different patient subj per study, n_t denotes the total number of patients per task, and y_t^i reflect whether the patient is diseased or non-diseased. For a given regression weight matrix $\boldsymbol{W} \in \mathbb{R}^{T \times d}$ for all the tasks, we denote $\boldsymbol{w}_t \in \mathbb{R}^d$ the row vector of $\boldsymbol{W}$ correspond to each task across features, and $\boldsymbol{w}_{(j)} \in \mathbb{R}^T$ the column vector of $\boldsymbol{W}$ corresponds to each feature across tasks. The Hadamard (element-wise) product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $\boldsymbol{a} \circ \boldsymbol{b}$, and diag denote the function map a vector to a diagonal matrix with the vector as its diagonal entries.

4.2.2 Hierarchy Bayesian Multitask Logistic Regression Model

Though the number of bacterial species are in the order of trillions on earth, it is well known that only a relative small number of bacteria are responsible for the majority of the bacterial infections in humans [180]. Hence our model assumes that only a few of the bacteria species are useful for the prediction task, where we impose sparsity on the regression coefficient through a Bernoulli-Gaussian distribution ($\boldsymbol{w}_t \circ \boldsymbol{z}$) [181, 156], where $\boldsymbol{w}_t$ control the magnitude of the effects and $\boldsymbol{z}$ controls the sparsity. This compound prior enforces some of the weights to be exactly zero, implying some of the bacteria species are irrelevant for predicting the health of patient. The sparsity term z are independently drawn from the same Bernoulli distribution with parameter θ. Note $\boldsymbol{z}$ does not depend on the specific task, this implies the sparsity pattern is shared across different tasks, which reflect the belief that there are few bacteria species are useful for prediction across all tasks. A hyper prior for θ is given by the beta distribution, which

utilize the conjugate property to control the overall sparsity level of the model. Further, to enforce the information sharing across different tasks, the row of $\boldsymbol{W}$, denoted by $\boldsymbol{w}_{(j)}$ are assumed to be *i.i.d* draws from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_0$. A Wishart prior is proposed for this shared covariance matrix to provide a way to utilized expert knowledge about the underling relationships among the studies, and this enables us to exploit the Wishart-normal conjugacy to obtain efficient inference later. The proposed conditional model can be summarized:

$$
\begin{aligned}
y_t^i|\boldsymbol{w}_t, \boldsymbol{z}; \boldsymbol{x}_t^i &\overset{\text{ind}}{\sim} \text{Bernoulli}\left(\text{sigmoid}\left(\left\langle \left(\boldsymbol{w}_t \circ \boldsymbol{z}\right), \boldsymbol{x}_t^i \right\rangle\right)\right) \quad \forall i = 1, \cdots n_t, \forall t = 1, \cdots T, \\
z_j|\theta &\overset{\text{i.i.d}}{\sim} \text{Bernoulli}\left(\theta\right) \quad \forall j = 1, \cdots d, \\
\theta &\sim \text{Beta}\left(\alpha_0, \beta_0\right), \\
\boldsymbol{w}_{(j)}|\boldsymbol{\Sigma}_0 &\overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_0\right) \quad \forall j = 1, \cdots d, \\
\boldsymbol{\Sigma}_0^{-1} &\sim \text{Wishart}\left(v_0, \boldsymbol{V}_0\right).
\end{aligned}
$$

where $\alpha_0, \beta_0, v_0, \boldsymbol{V}_0$ are hyperparameters selected by the experimenter. Smaller value of the ratio $\frac{\alpha_0}{\alpha_0+\beta_0}$ corresponds to a prior belief of fewer informative features while the magnitude α_0 controls the confidence of prior belief, and $v_0, \boldsymbol{V}_0$ reflects the prior knowledge about the covariance structure of the regression coefficient $\boldsymbol{W}$ across different tasks.

The proposed model leads to a log conditional probability, up to an unimpor-

tant constant:

$$
\begin{aligned}
\log\left(p\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\Sigma}_0, \theta \mid \boldsymbol{X}; v_0, \boldsymbol{V}_0, \alpha_0, \beta_0\right)\right) = & -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{V}_0^{-1}\boldsymbol{\Sigma}_0^{-1}\right) + \frac{v_0 + d - T - 1}{2}\log\det\left(\boldsymbol{\Sigma}_0^{-1}\right) \\
& - \frac{v_0}{2}\log\det\left(\boldsymbol{V}_0\right) + \left(\alpha_0 - 1\right)\log\left(\theta\right) + \left(\beta_0 - 1\right)\log\left(1 - \theta\right) \\
& + \log\Gamma\left(\alpha_0 + \beta_0\right) - \log\Gamma\left(\alpha_0\right) - \log\Gamma\left(\beta_0\right) \\
& - \sum_j \frac{1}{2}\left\langle \boldsymbol{w}_{(j)}, \boldsymbol{\Sigma}_0^{-1}\boldsymbol{w}_{(j)} \right\rangle + \sum_t \sum_i y_t^i \left(\left\langle \left(\boldsymbol{w}_t \circ \boldsymbol{z}\right), \boldsymbol{x}_t^i \right\rangle\right) \\
& - \sum_t \sum_i \log\left(\exp\left(\left\langle \left(\boldsymbol{w}_t \circ \boldsymbol{z}\right), \boldsymbol{x}_t^i \right\rangle\right) + 1\right) \\
& + \left(\sum_j z_j\right)\log\theta + \left(d - \sum_j z_j\right)\log\left(1 - \theta\right). \qquad (4.1)
\end{aligned}
$$

where Γ denote the gamma function.

4.3 Variational Inference

With the combination of the logistic function and the hierarchical structure, inference from the exact posterior distribution of the conditional model is intractable since the posterior distribution is not available in closed form. We resort to a variational approach [55, 157] where we approximate the posterior distribution with a simpler distribution, and the approximation is iteratively refined. We refer interested reader to [55] for a comprehensive review on variational inference (VI) as a general approach for Bayesian inference.

Section 4.3.1 introduces the mean-field approximation used to approximate the posterior distribution along with the variational objective function, and Section 4.3.2 summarize the optimization algorithm based on coordinate ascent [57].

4.3.1 Mean-Field Approximation and Variational Lower Bound

Mean field approximation is a prevalent choice of the approximation family, because it is expressive enough to approximate the complex posteriors, and simple

enough to lead to tractable computations [57].

In this chapter, we propose to approximate the true posterior of the proposed model with:

$$q(\theta, \boldsymbol{W}, \boldsymbol{\Sigma}_0, \boldsymbol{z}) = q\left(\theta; \alpha, \beta\right) q\left(\boldsymbol{\Sigma}_0^{-1}; v, \boldsymbol{V}\right) \prod_j q\left(z_j; \phi_j\right) q(\boldsymbol{w}_{(j)}; \boldsymbol{m}_{(j)}, \boldsymbol{\Sigma}_j). \tag{4.2}$$

Where $q\left(\theta; \alpha, \beta\right)$ is a beta distribution with parameters (α, β), $q\left(z_j; \phi_j\right)$ is a Bernoulli distribution with parameters ϕ_i, $q\left(\boldsymbol{\Sigma}_0^{-1}; v, \boldsymbol{V}\right)$ is a Wishart distribution with parameters $(v, \boldsymbol{V})$, and $q\left(\boldsymbol{w}_{(j)}; \boldsymbol{m}_{(j)}, \boldsymbol{\Sigma}_j\right)$ is a multivariate Gaussian distribution with mean $\boldsymbol{m}_{(j)}$ and covariance $\boldsymbol{\Sigma}_j$ for $j = 1, \ldots, d$.

The optimization objective of variational inference (VI) is to minimize the KL-divergence between the approximation Eqn. 4.2 and the true posterior by maximize evidence lower bound(ELBO):

$$\begin{aligned}
\operatorname{ELBO}(\boldsymbol{q}) &= \mathbb{E}_q\left[\log f\left(\boldsymbol{X}, \boldsymbol{Y}, \theta, \boldsymbol{z}, \boldsymbol{W}, \boldsymbol{\Sigma}_0\right)\right] + \operatorname{Entropy}(q) \\
&= -\frac{1}{2} \operatorname{tr}\left(v \boldsymbol{V}_0^{-1} \boldsymbol{V}\right) + \frac{v_0 + d}{2} \log\det\left(\boldsymbol{V}\right) \\
&+ \frac{v_0 + d - v}{2} \psi_T\left(\frac{v}{2}\right) + \frac{vT}{2} + \ln \Gamma_T\left(\frac{v}{2}\right) \\
&+ \left(\alpha_0 + \sum_j \phi_j - \alpha\right) \psi(\alpha) + \left(\beta_0 + d - \sum_j \phi_j - \beta\right) \psi(\beta) \\
&+ \ln B(\alpha, \beta) + (\alpha + \beta - d - \alpha_0 - \beta_0) \psi(\alpha + \beta) \\
&- \frac{1}{2} \sum_j \operatorname{tr}\left(v \boldsymbol{V}\left(\boldsymbol{m}_{(j)} \boldsymbol{m}_{(j)}^{\top} + \boldsymbol{\Sigma}_j\right)\right) \\
&+ \sum_t \sum_i y_t^i \left(\left\langle \left(\boldsymbol{m}_t \circ \boldsymbol{\phi}\right), \boldsymbol{x}_t^i \right\rangle\right) \\
&- \sum_t \sum_i \mathbb{E}_{W, z \sim q}\left[\log\left(\exp\left(\left\langle \left(\boldsymbol{w}_t \circ \boldsymbol{z}\right), \boldsymbol{x}_t^i \right\rangle\right) + 1\right)\right] \\
&+ \frac{1}{2} \sum_j \log\det\left(\boldsymbol{\Sigma}_j\right) - \sum_j \phi_j \log\left(\phi_j\right) - \sum_j \left(1 - \phi_j\right) \log\left(1 - \phi_j\right).
\end{aligned} \tag{4.3}$$

where ψ is the digamma function, ψ_T is the multivariate digamma function

and γ_T is the multivariate gamma function. Since the expectations of the sigmoid functions do not admit closed form solutions, we approximate the sigmoid functions by quadratic lower bounds:

$$
\begin{aligned}
-\log\left(\exp\left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right)+1\right) \geq & -\log\left(\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right)+1\right) \\
& - \frac{\left\langle \boldsymbol{x}_t^i, \boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}' \right\rangle}{\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right)+1} \\
& - \frac{1}{8}\left(\boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}'\right)^\top \boldsymbol{x}_t^i \left(\boldsymbol{x}_t^i\right)^\top \left(\boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}'\right).
\end{aligned} \tag{4.4}
$$

for $i = 1, \ldots, n_t, t = 1, \ldots, T$, and $\boldsymbol{w}_t', \boldsymbol{z}'$ are deterministic reference points of choice. This type of approximations has been used to design majorizor-minimization (MM) algorithms for the logistic regression problem [182, 183]. The resulting approximation is, up to a constant:

$$
\begin{aligned}
-\sum_t \sum_i \mathbb{E}_{\boldsymbol{W}, \boldsymbol{z} \sim q}\left[\log\left(\exp\left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right)+1\right)\right] \geq & -\sum_t \sum_i \log\left(\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right)+1\right) \\
& + \sum_t \sum_i \frac{\left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle}{\exp\left(-\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right)+1} \\
& - \frac{1}{8}\sum_t \sum_i \left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle^2 \\
& - \sum_t \sum_i \frac{1}{\exp\left(-\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right)+1} \left\langle \boldsymbol{x}_t^i, (\boldsymbol{m}_t \circ \boldsymbol{\phi}) \right\rangle \\
& + \frac{1}{4}\sum_t \sum_i \left(\left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle\right)\left(\left\langle \boldsymbol{m}_t \circ \boldsymbol{\phi}, \boldsymbol{x}_t^i \right\rangle\right) \\
& - \frac{1}{8}\sum_t \sum_i \left(\left\langle \boldsymbol{m}_t \circ \boldsymbol{\phi}, \boldsymbol{x}_t^i \right\rangle\right)^2 \\
& + \frac{1}{8}\sum_t \sum_j \left(m_{t,j}^2 \left(\phi_j - 1\right) - \left(\boldsymbol{\Sigma}_j\right)_{t,t}\right)\phi_j \sum_i \left(x_t^{ij}\right)^2.
\end{aligned} \tag{4.5}
$$

There are other alternative approaches using different lower bounds [184, 185], chapter 10.6 of [57], but they require additional variational parameters scale with the feature dimension (d), which complicates the variational computations.

4.3.2 Coordinate Ascent Variational Inference (CAVI)

Coordinate ascent variational inference (CAVI) [57] is a optimization technique where we optimize one set of the variational parameters at a time while holding the others fixed.

One useful result (Equation 18 of [55]) states that if we are to approximate a general posterior distribution $p\left(\boldsymbol{\xi} \mid \text{data}\right)$ with a mean-field approximation $q\left(\boldsymbol{\xi}\right) := \prod_j q_j\left(\xi_j\right)$, the CAVI update for j-th latent variable ξ_j (i.e the optimal solution $q_j^\star\left(\xi_j\right)$) is proportional to the exponentiated conditional expected log of the joint:

$$q_j^\star \propto \exp\left(\mathbb{E}_{\boldsymbol{\xi}_{-j} \sim q_{-j}}\left[\log\left(p\left(\xi_j, \boldsymbol{\xi}_{-j} \mid \text{data}\right)\right)\right]\right). \tag{4.6}$$

where $\boldsymbol{\xi}_{-j}$ corresponds to all but the j-th latent variable.

The resulting Coordinate Ascent Variational Inference (CAVI) algorithm is summarized in Algorithm. 2, and see section 4.6.1 for details of the derivations using Eqn. 4.6.

Algorithm 2 CAVI for Bayesian Multitask Sparse Logistic Regression

procedure CAVI(Inputs: $\left(\boldsymbol{x}_t^i, y_t^i\right)_{t=1,\cdots T, i=1,\cdots n_t}$)
  for all itr $= 1, \cdots,$ Niter **do**
    for all $j = 1, \cdots d$ **do**
      $\boldsymbol{\Sigma}_j \leftarrow$ by Eqn. 4.7
    end for
    for all $j = 1, \cdots, d$ **do**
      $\boldsymbol{m}_{(j)} \leftarrow$ by Eqn. 4.9
    end for
    for all $j = 1, \cdots, d$ **do**
      $\phi_j \leftarrow$ by Eqn. 4.11
    end for
    $\alpha \leftarrow \alpha_0 + \sum_j \phi_j$,
    $\beta \leftarrow \beta_0 + d - \sum_j \phi_j$,
    $\theta \leftarrow \frac{\alpha}{\alpha+\beta}$,
    $v \leftarrow v_0 + d$,
    $\boldsymbol{V} \leftarrow \left(\boldsymbol{V}_0^{-1} + \sum_j \boldsymbol{m}_{(j)} \boldsymbol{m}_{(j)}^\top + \boldsymbol{\Sigma}_j\right)^{-1}$
  end for
end procedure

4.4 Experiments

In this section, we evaluate the performance of our proposed method on both simulated data and real microbiome data pooled from multiple studies [158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176], and we compare it with following methods:

- Single-task Logistic Classifier with sparsity penalty (STL-LC) [178, 179]: this is a single-task model, where we fit independent logistic models to each task separately. This is an extension of standard LASSO into the binary classification setting.
- Pooled Logistic Classifier with sparsity penalty (Pooled-LC): we train a single logistic regression model for all tasks.
- MTFL (Multitask Feature Learning) [149]: this is an optimization based approach to multitask learning based on $\ell_{2,1}$-norm regularization. The proposed method can be seen as a Bayesian Hierarchical extension, and the difference in modeling is how the sparsity pattern is encouraged: we use a overparameterized Bernoulli-Gaussian model, which has better support recovery properties [156].
- MSSL (Multitask Sparsity Structure Learning) [151]: this is an optimization based multitask learning approach, where the imposed sparsity structure is on the precision (inverse covariance matrix) of the regression coefficients across tasks. The optimization problem of this formulation is equivalent to the graphical lasso problem [186, 65] for covariance estimation.

Model Selections: For all the methods, we employ a model selection strategy based on cross-entropy loss on the validation dataset to select the hyperparameters (e.g $\alpha_0, \beta_0, v_0, \boldsymbol{V}_0$ of the proposed Bayesian approach). Specifically, we use 10 repeated runs of stratified cross-validation [187] since we have limited number of samples and class labels are imbalanced.

Evaluation Metrics: the microbiome dataset that we used has imbalanced class labels, i.e there are more healthy patients than diseased patients. Thus we use following metrics to quantify the performance of predictive models: accuracy, balanced accuracy, averaged precision, F1 score, F2 score, Matthews correlation coefficient (MCC), and area under the curve (AUC). Table 4.1 summarizes the definitions of these metrics.

Table 4.1 Definitions of classification metrics and the intermediate variables given ground truth labels $\boldsymbol{y}$ and predicted labels $\hat{\boldsymbol{y}}$. $\wedge$ denotes the "and" operation, and $\mathbf{1}(\cdot)$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise.

Metric/Immediate Variable	**Definition**
TP (True Positives)	$\text{TP} = \sum_{i=1}^{N} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 1)$
TN (True Negatives)	$\text{TN} = \sum_{i=1}^{N} \mathbf{1}(y_i = 0 \wedge \hat{y}_i = 0)$
FP (False Positives)	$\text{FP} = \sum_{i=1}^{N} \mathbf{1}(y_i = 0 \wedge \hat{y}_i = 1)$
FN (False Negatives)	$\text{FN} = \sum_{i=1}^{N} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 0)$
Precision	$\frac{\text{TP}}{\text{TP+FP}}$
Recall	$\frac{\text{TP}}{\text{TP+FN}}$
Accuracy	$\text{Accuracy}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\text{TP+TN}}{N}$
Balanced Accuracy	$\text{Balanced Accuracy}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$
Average Precision	Integral of precision over all recall levels
F1 Score	$\text{F1 Score}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$
F2 Score	$\text{F2 Score}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 5 \times \frac{\text{Precision} \times \text{Recall}}{4 \times \text{Precision+Recall}}$
MCC	$\text{MCC}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}$
AUC	Area under the ROC curve plotting TP rates vs. FP rates

4.4.1 Synthetic Datasets

We generate synthetic datasets to examine whether our algorithm is able to: 1) recover the support of the regression coefficients that correspond to informative features for the prediction, which are evaluated by the metrics from Table 4.1; 2) recover the ground truth regression coefficients (up to normalization), which are evaluated by cosine distance (i.e one minus the normalized inner product).

To mimic various characteristics of a real world dataset, we generate six datasets with varying sparsity level (the common support of regression coefficients across tasks) and class imbalance (whether sample sizes across different tasks are of the same magnitude). Additional details of each dataset are summarized in Table 4.2.

Support Recovery: We evaluate the support recovery of the algorithms by turning the support recovery problem into a binary prediction problem. The result is summarized in Table 4.3. The proposed Bayesian method is the best performing algorithm in most of the metrics across all settings. In particular, when the ground truth regression coefficients have a sparse support, the proposed method has a close to perfect recovery (98.8%).

Weight Recovery: We evaluate the prediction performance of the algorithm by assess how well the algorithms can recover the ground truth regression coefficients. Since the logistic prediction is scale invariant, we evaluate the results by the cosine distance. The result is summarized in Table 4.4. The proposed Bayesian method is the best performing algorithm in all but the dense case.

Table 4.2 Summary of the simulated dataset, where $\sim$ Pois means the number of samples is Poisson distributed, and $\sim$ NB means the number of sample follow a negative binomial distribution. For all the simulation the number of features is 100 and number of tasks is 10. For the unbalanced datasets, We add all the sample sizes by 6 to ensure that both positive samples and negative samples are present across all tasks. Both settings have an expected sample size 30, with the imbalanced case has more variations of sample sizes among different tasks. The θ parameter corresponds to the expected percentage of the predictive features.

	Dense ($\theta = 0.8$)	Sparse ($\theta = 0.2$)	Ultra Sparse ($\theta = 0.05$)
Balanced $\sim$ Pois $(24) + 6$	dataset1	dataset2	dataset3
Unbalanced $\sim$ NB $(1, 0.04) + 6$	datasett4	dataset5	dataset6

Table 4.3 Summary of the support recovery results for the simulated data. The bold number means the corresponding method is the best performing algorithm for the given metrics and dataset, and the values in parentheses represent standard deviations computed over 10 different runs. The proposed Bayesian approach outperforms the benchmark methods in all evaluation metrics when there is a shared sparsity structure across regression coefficients of different tasks. Both MSSL and MTFL prioritize the prediction performance in the cross-validation step which results in complete dense solutions (i.e all regression coefficients are non-zero), hence they have identical results.

Dataset	Metrics	BayesMTL	MTFL	MSSL	STL-LC	Pooled-LC
dataset1 (dense, balanced)	Accuracy	0.322 (0.06)	**0.793** (0.05)	**0.793** (0.05)	0.688 (0.05)	0.71 (0.04)
	Balanced Accuracy	**0.565** (0.03)	0.5 (0)	0.5 (0)	0.525 (0.05)	0.52 (0.03)
	Average Precision	**0.819** (0.05)	0.793 (0.05)	0.793 (0.05)	0.803 (0.04)	0.800 (0.05)
	F1 Score	0.261 (0.06)	**0.884** (0.03)	**0.884** (0.03)	0.801 (0.04)	0.821 (0.03)
	F2 Score	0.183 (0.05)	**0.950** (0.01)	**0.950** (0.01)	0.800 (0.05)	0.834 (0.03)
	AUC	**0.565** (0.03)	0.5 (0)	0.5 (0)	0.525 (0.05)	0.520 (0.03)
	MCC	**0.156** (0.06)	0 (0)	0 (0)	0.0565 (0.09)	0.0427 (0.07)
dataset2 (sparse, balanced)	Accuracy	**0.882** (0.04)	0.225 (0.03)	0.225 (0.03)	0.438 (0.08)	0.364 (0.02)
	Balanced Accuracy	**0.768** (0.05)	0.5 (0)	0.5 (0)	0.616 (0.05)	0.548 (0.04)
	Average Precision	**0.597** (0.10)	0.225 (0.03)	0.225 (0.03)	0.279 (0.06)	0.244 (0.02)
	F1 Score	**0.681** (0.09)	0.367 (0.03)	0.367 (0.03)	0.431 (0.07)	0.383 (0.03)
	F2 Score	0.602 (0.09)	0.590 (0.03)	0.590 (0.03)	**0.634** (0.06)	0.579 (0.04)
	AUC	**0.768** (0.05)	0.5 (0)	0.5 (0)	0.616 (0.05)	0.548 (0.04)
	MCC	**0.638** (0.10)	0 (0)	0 (0)	0.227 (0.09)	0.104 (0.07)
dataset3 (ultra sparse, balanced)	Accuracy	**0.988** (0.02)	0.0350 (0.01)	0.0350 (0.01)	0.437 (0.08)	0.199 (0.04)
	Balanced Accuracy	**0.947** (0.06)	0.5 (0)	0.5 (0)	0.708 (0.04)	0.526 (0.09)
	Average Precision	**0.801** (0.20)	0.035 (0.01)	0.035 (0.01)	0.0584 (0.01)	0.0372 (0.01)
	F1 Score	**0.876** (0.13)	0.0674 (0.02)	0.0674 (0.02)	0.110 (0.02)	0.070 (0.02)
	F2 Score	**0.886** (0.12)	0.152 (0.04)	0.152 (0.04)	0.235 (0.05)	0.154 (0.04)
	AUC	**0.947** (0.06)	0.5 (0)	0.5 (0)	0.708 (0.04)	0.526 (0.09)
	MCC	**0.879** (0.13)	0 (0)	0 (0)	0.154 (0.02)	0.0199 (0.08)
dataset4 (dense, unbalanced)	Accuracy	0.441 (0.12)	**0.795** (0.04)	**0.795** (0.04)	0.689 (0.09)	0.693 (0.04)
	Balanced Accuracy	**0.584** (0.06)	0.5 (0)	0.5 (0)	0.554 (0.07)	0.517 (0.04)
	Average Precision	**0.827** (0.04)	0.795 (0.04)	0.795 (0.04)	0.814 (0.04)	0.801 (0.04)
	F1 Score	0.483 (0.16)	**0.885** (0.02)	**0.885** (0.02)	0.794 (0.08)	0.808 (0.03)
	F2 Score	0.393 (0.15)	**0.951** (0.01)	**0.951** (0.01)	0.788 (0.12)	0.813 (0.04)
	AUC	**0.584** (0.06)	0.5 (0)	0.5 (0)	0.554 (0.07)	0.517 (0.04)
	MCC	**0.147** (0.09)	0 (0)	0 (0)	0.118 (0.12)	0.039 (0.08)
dataset5 (sparse, unbalanced)	Accuracy	**0.796** (0.07)	0.219 (0.04)	0.219 (0.04)	0.451 (0.07)	0.334 (0.03)
	Balanced Accuracy	**0.783** (0.06)	0.5 (0)	0.5 (0)	0.626 (0.036)	0.541 (0.04)
	Average Precision	**0.472** (0.13)	0.219 (0.04)	0.219 (0.04)	0.275 (0.05)	0.234 (0.03)
	F1 Score	**0.625** (0.1)	0.358 (0.05)	0.358 (0.05)	0.427 (0.06)	0.370 (0.04)
	F2 Score	**0.694** (0.05)	0.579 (0.06)	0.579 (0.06)	0.630 (0.05)	0.569 (0.04)
	AUC	**0.783** (0.06)	0.5 (0)	0.5 (0)	0.626 (0.04)	0.541 (0.04)
	MCC	**0.516** (0.13)	0 (0)	0 (0)	0.240 (0.05)	0.086 (0.08)
dataset6 (ultra sparse, unbalanced)	Accuracy	**0.917** (0.10)	0.059 (0.0239)	0.059 (0.0239)	0.478 (0.07)	0.21 (0.04)
	Balanced Accuracy	**0.905** (0.08)	0.5 (0)	0.5 (0)	0.715 (0.05)	0.553 (0.03)
	Average Precision	**0.598** (0.02)	0.059 (0.02)	0.059 (0.02)	0.097 (0.03)	0.0646 (0.02)
	F1 Score	**0.699** (0.27)	0.11 (0.04)	0.11 (0.04)	0.176 (0.05)	0.120 (0.04)
	F2 Score	**0.763** (0.20)	0.233 (0.08)	0.233 (0.08)	0.343 (0.08)	0.246 (0.07)
	AUC	**0.905** (0.08)	0.5 (0)	0.5 (0)	0.715 (0.05)	0.553 (0.03)
	MCC	**0.704** (0.26)	0 (0)	0 (0)	0.198 (0.03)	0.063 (0.03)

Table 4.4 Summary of the weights recovery measured in cosine distance. The cosine distance is bounded between 0 and 2 with 0 means perfect recovery. The proposed Bayesian approach outperforms the benchmark methods in all evaluation metrics when there is a shared sparsity structure across regression coefficients of different tasks.

Dataset	BayesMTL	MTFL	MSSL	STL-LC	Pooled-LC
dataset1 (dense, balanced)	0.736 (0.06)	0.561 (0.02)	**0.554** (0.02)	0.735 (0.02)	0.983 (0.04)
dataset2 (sparse, balanced)	**0.341** (0.09)	0.566 (0.01)	0.563 (0.01)	0.560 (0.05)	1.02 (0.05)
dataset3 (ultra sparse, balanced)	**0.098** (0.04)	0.589 (0.03)	0.598 (0.02)	0.257 (0.04)	0.995 (0.12)
dataset4 (dense, unbalanced)	0.670 (0.10)	0.582 (0.08)	**0.576** (0.05)	0.736 (0.07)	1.02 (0.03)
dataset5 (sparse, unbalanced)	**0.456** (0.11)	0.597 (0.08)	0.576 (0.04)	1.04 (0.06)	0.571 (0.09)
dataset6 (ultra sparse, unbalanced)	**0.223** (0.12)	0.651 (0.05)	0.612 (0.04)	0.353 (0.11)	1.07 (0.10)

4.4.2 Microbiome Data

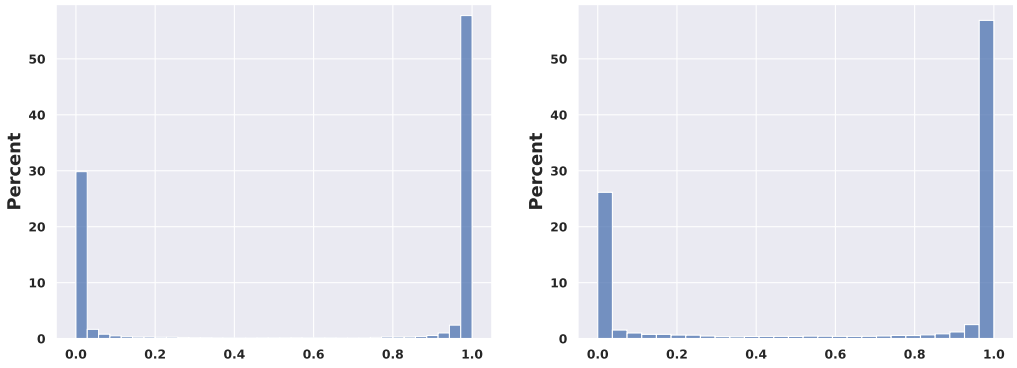
Our goal is two folds: 1) show that multitask learning on the one-hot coded vector of diseases can perform the disease classification with uncertainty quantification 2) identify the common bacteria that are most predictive for a patient's health.

Prediction Performance: analogous to previous subsection, we evaluate the prediction performance through various metrics from Table 4.1, and the results are summarized in Table 4.5. Due to the heterogeneous nature of the data (i.e pooled

from studies with different experimental objectives, laboratory setups, sequencing equipments, patient demographics etc.), we do not see an improvement of the multitask learning models. However, the proposed model is the most consistent approach that provide sparse solutions.

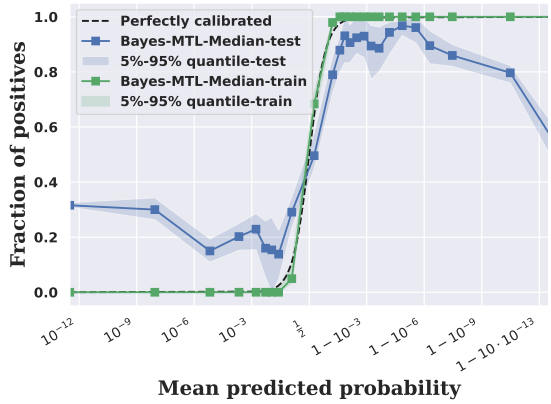
Feature Importance: we assign feature importance to each of the OTUs by combining the magnitude of regression coefficient ($\{\boldsymbol{w}_t\}$) and the sparsity parameters (z). From the estimated posterior distribution, we draw samples to explore the full distribution of the feature importance. The most important features will corresponds to the features with consistent high importance weights across draws. The result is summarized in Fig. 4.2. The proposed model learns a sparse set of features shared across different datasets from the data as reflected by colored strips.

Goodness of fit: We evaluate the goodness of fit of the proposed method through calibration curve [188], which plots the predicted probability against the observed labels. For a well calibrated probabilistic model, among all the samples model predicted with probability p% being healthy, close to p% of them will indeed be healthy. The calibration results are summarized in Fig. 4.1 for both the training data and test data. The proposed Bayesian approach provides additional uncertainty quantification about the predicted probabilities. Since the proposed model is probabilistic, it provides well calibrated results. The performance degrades at the boundary values for the test data, which indicate the choice of logit function as a link function is resulting in over-confident predictions. We discuss possible extensions in Section 4.5.



(a) Predicted probabilities on training data.

(b) Predicted probabilities on test data.

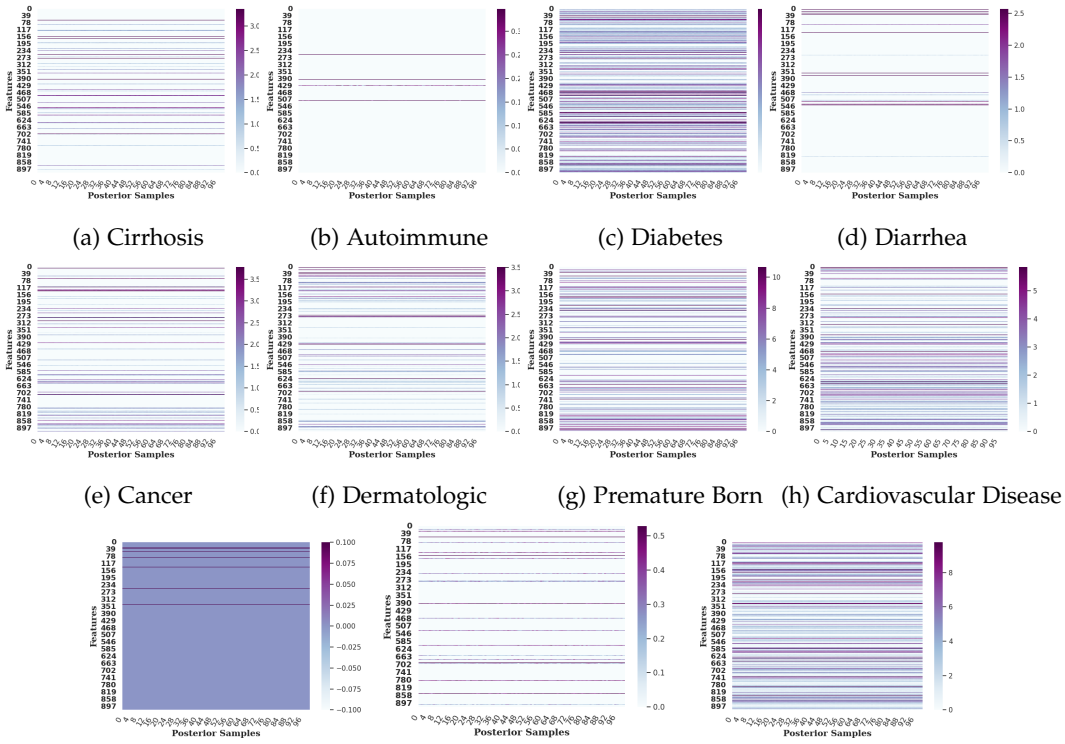


(c) Calibration curves for the training data and the test data with 90% marginal credible intervals.

Figure 4.1: Calibration analysis for the proposed model on the Order Taxon level. Fig. (a) and Fig. (b) show the histograms of the predicted probabilities and training and test data respectively. Due the choice of logit as link function, the predicted probabilities are concentrated around the boundaries. Fig. (c) show the calibration curves of the predictions from training and test data. The model achieves near perfect calibration on the training data, and the degradation of performance on the test data at the boundary values indicates that the logit function as a link function is resulting in over-confident predictions.

Table 4.5 Summary of the prediction performance. The bold number means the corresponding method is the best performing algorithm for the given metrics and taxon level, and the values in parentheses represent standard deviations computed over 5 different runs. Due to the heterogeneous nature of the data, we do not see an improvement of the proposed approach over single-tasked model. However, the proposed approach is the only multitask method that provides a sparse solution i.e identify common bacteria across studies of the same disease category that are informative for the predictions.

Taxon Levels	Metrics	BayesMTL	MTFL	MSSL	Pooled-LC	STL-LC
Kingdom	Accuracy	0.679 (0.0131)	0.695 (0.0145)	**0.699** (0.0198)	0.582 (0.0181)	0.63 (0.021)
	Balanced Accuracy	0.602 (0.012)	0.619 (0.0114)	**0.64** (0.0165)	0.54 (0.0184)	0.619 (0.0117)
	Average Precision	0.636 (0.00747)	0.645 (0.00563)	**0.657** (0.00901)	0.585 (0.00857)	0.635 (0.0103)
	F1 Score	0.674 (0.0253)	0.674 (0.013)	0.69 (0.023)	**0.698** (0.0112)	0.576 (0.036)
	F2 Score	0.696 (0.0244)	0.688 (0.0172)	0.701 (0.0272)	**0.807** (0.0126)	0.591 (0.0467)
	AUC	0.602 (0.012)	0.619 (0.0114)	0.64 (0.0165)	0.54 (0.0184)	**0.619** (0.0117)
	MCC	0.222 (0.0253)	0.255 (0.0294)	**0.296** (0.0466)	0.104 (0.0446)	0.229 (0.0264)
	Sparsity Ratio	**0.246** (0.0315)	0.534 (0.0124)	0.542 (0.00843)	0.982 (0.0364)	0.271 (0.0406)
Phylum	Accuracy	0.699 (0.0296)	**0.722** (0.0132)	0.677 (0.00977)	0.582 (0.0304)	0.655 (0.0211)
	Balanced Accuracy	0.647 (0.0286)	**0.696** (0.0161)	0.674 (0.0164)	0.563 (0.0303)	0.631 (0.0141)
	Average Precision	0.666 (0.0206)	**0.695** (0.0137)	0.673 (0.00766)	0.599 (0.0157)	0.651 (0.0124)
	F1 Score	0.713 (0.0333)	**0.733** (0.0145)	0.68 (0.00808)	0.664 (0.0207)	0.595 (0.0555)
	F2 Score	0.732 (0.0308)	**0.739** (0.0136)	0.674 (0.0103)	0.725 (0.0222)	0.611 (0.0609)
	AUC	0.647 (0.0286)	**0.696** (0.0161)	0.674 (0.0164)	0.563 (0.0303)	0.631 (0.0141)
	MCC	0.309 (0.0563)	**0.396** (0.0338)	0.335 (0.0264)	0.134 (0.0647)	0.266 (0.0273)
	Sparsity Ratio	0.215 (0.0518)	0.492 (0.0324)	0.564 (0.0159)	0.883 (0.0254)	**0.114** (0.0219)
Class	Accuracy	0.703 (0.0168)	0.679 (0.0263)	0.645 (0.0133)	0.624 (0.00985)	**0.717** (0.0118)
	Balanced Accuracy	0.642 (0.0163)	**0.68** (0.0282)	0.652 (0.0179)	0.612 (0.0121)	0.677 (0.0179)
	Average Precision	0.655 (0.00783)	0.682 (0.0152)	0.658 (0.00646)	0.624 (0.00459)	**0.683** (0.0127)
	F1 Score	**0.699** (0.0172)	0.68 (0.0276)	0.655 (0.016)	0.689 (0.00774)	0.671 (0.0359)
	F2 Score	0.725 (0.0206)	0.681 (0.0266)	0.649 (0.0167)	**0.743** (0.0143)	0.68 (0.037)
	AUC	0.642 (0.0163)	**0.68** (0.0282)	0.652 (0.0179)	0.612 (0.0121)	0.677 (0.0179)
	MCC	0.3 (0.0327)	0.351 (0.0521)	0.291 (0.0315)	0.236 (0.0278)	**0.354** (0.0387)
	Sparsity Ratio	0.179 (0.0362)	0.458 (0.029)	0.572 (0.0154)	0.847 (0.0214)	**0.0963** (0.0164)
Order	Accuracy	0.699 (0.0242)	0.628 (0.0978)	0.625 (0.0176)	0.622 (0.0151)	**0.709** (0.0205)
	Balanced Accuracy	0.653 (0.028)	0.641 (0.0731)	0.636 (0.0183)	0.606 (0.0113)	**0.669** (0.0167)
	Average Precision	0.66 (0.0141)	0.655 (0.0459)	0.648 (0.00811)	0.625 (0.00612)	**0.679** (0.00903)
	F1 Score	**0.703** (0.023)	0.544 (0.272)	0.641 (0.0131)	0.678 (0.0197)	0.666 (0.0401
	F2 Score	**0.725** (0.0246)	0.543 (0.272)	0.638 (0.0103)	0.723 (0.0286)	0.681 (0.0456)
	AUC	0.653 (0.028)	0.641 (0.0731)	0.636 (0.0183)	0.606 (0.0113)	**0.669** (0.0167)
	MCC	0.318 (0.0658)	0.276 (0.143)	0.26 (0.0312)	0.217 (0.026)	**0.342** (0.038)
	Sparsity Ratio	0.132 (0.00943)	0.462 (0.236)	0.592 (0.0188)	0.833 (0.059)	**0.102** (0.0576)
Family	Accuracy	0.683 (0.0121)	0.651 (0.00838)	0.64 (0.0143)	0.634 (0.0211)	**0.704** (0.0384)
	Balanced Accuracy	0.637 (0.0103)	0.654 (0.013)	0.649 (0.014)	0.62 (0.0242)	**0.691** (0.0261)
	Average Precision	0.655 (0.00972)	0.663 (0.00527)	0.656 (0.00786)	0.634 (0.0156)	**0.694** (0.0175)
	F1 Score	0.682 (0.0182)	0.635 (0.0434)	0.651 (0.0112)	**0.688** (0.0174)	0.636 (0.062))
	F2 Score	0.698 (0.0193)	0.636 (0.0505)	0.646 (0.0114)	**0.731** (0.0184)	0.646 (0.061)
	AUC	0.637 (0.0103)	0.654 (0.013)	0.649 (0.014)	0.62 (0.0242)	**0.691** (0.0261)
	MCC	0.279 (0.0177)	0.301 (0.0278)	0.286 (0.0248)	0.244 (0.0462)	**0.383** (0.0528)
	Sparsity Ratio	**0.0974** (0.0442)	0.539 (0.0588)	0.611 (0.0256)	0.783 (0.115)	0.106 (0.0597)
Genus	Accuracy	0.697 (0.01)	0.662 (0.0147)	0.658 (0.0111)	0.656 (0.0121)	**0.769** (0.0104)
	Balanced Accuracy	0.663 (0.00891)	0.665 (0.0174)	0.662 (0.0131)	0.652 (0.0199)	**0.693** (0.0164)
	Average Precision	0.671 (0.00646)	0.67 (0.00739)	0.669 (0.00666)	0.654 (0.0131)	**0.7** (0.0141)
	F1 Score	**0.706** (0.00669)	0.656 (0.0379)	0.665 (0.011)	0.695 (0.00896)	0.674 (0.0491)
	F2 Score	**0.726** (0.0138)	0.652 (0.0392)	0.656 (0.00992)	0.726 (0.00819)	0.688 (0.0483)
	AUC	0.663 (0.00891)	0.665 (0.0174)	0.662 (0.0131)	0.652 (0.0199)	**0.693** (0.0164)
	MCC	0.331 (0.0193)	0.321 (0.0352)	0.313 (0.0246)	0.3 (0.039)	**0.385** (0.0301)
	Sparsity Ratio	**0.0956** (0.0799)	0.467 (0.0931)	0.621 (0.0185)	0.894 (0.119)	0.136 (0.0376)
Species	Accuracy	0.693 (0.0159)	0.661 (0.022)	0.709 (0.0195)	0.669 (0.0135)	**0.725** (0.0316)
	Balanced Accuracy	0.673 (0.0165)	0.664 (0.0202)	0.701 (0.0211)	0.667 (0.0104)	**0.705** (0.0162)
	Average Precision	0.676 (0.00819)	0.671 (0.0134)	0.705 (0.013)	0.669 (0.00927)	**0.716** (0.00852)
	F1 Score	0.695 (0.0183)	0.648 (0.0715)	**0.705** (0.0205)	0.691 (0.0144)	0.694 (0.0366)
	F2 Score	0.704 (0.0241)	0.651 (0.0735)	0.698 (0.0225)	0.702 (0.0205)	**0.71** (0.0454)
	AUC	0.673 (0.0165)	0.664 (0.0202)	0.701 (0.0211)	0.667 (0.0104)	**0.705** (0.0162)
	MCC	0.348 (0.0286)	0.324 (0.0363)	0.395 (0.0386)	0.333 (0.0241)	**0.41** (0.0342)
	Sparsity Ratio	**0.134** (0.077)	0.638 (0.0789)	0.579 (0.0285)	0.797 (0.142)	0.194 (0.0865)



(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.2: Feature importance weight visualization across 11 different disease category of Order taxon level. The x-axis corresponds to different samples draw from the posterior distribution and the y-axis correspond to different OTUs. The gradation from white to black for a variable's color corresponds to its increasing importance weight, and the darker shaded horizontal lines represent the sparse features selected by the algorithm.

4.5 Conclusion

In this chapter a hierarchical Bayesian multitask logistic regression model is proposed to perform human healthy conditions from multiple related human gut bacteria abundance data. The model is designed to select common informative features across different tasks through the built-in sparsity structure. We derive a computationally efficient inference algorithm based on variational inference. Our simulation studies show that the proposed approach excels in situations when there are shared sparsity structures of the regression coefficients across the different tasks. Our experiments on a real world dataset pooled from multiple

studies demonstrate the utility of the method to extract informative taxons while providing well-calibrated predictions with uncertainty quantification.

There are several directions for future work. One direction is to replace the logit function with other link functions (e.g a probit link function) that have flatter tails so the model is less prone to overconfidence. Second direction is to extend our model to multi-label classification problems, where each task contains multiple binary predictions (e.g diagnosis of different diseases on the same patient). This generalization is of particular interests to the human health prediction application considered in this chapter, since the diseases are not mutually exclusive. Another related extension is to consider the multiclass classification problem, where each task is a classification problem with more than 2 labels (e.g different stages of a disease).

4.6 Appendix

4.6.1 CAVI update derivation

This subsection includes the derivations of CAVI updates for Algorithm. 2.

Update for α, β: Based on Eqn. 4.6, the exponentiated conditional expectation of all the parameters except θ up to a constant scaling factor:

$$q^\star(\theta) \propto \exp\left((\alpha_0 - 1)\log\theta + (\beta_0 - 1)\log(1-\theta) + \left(\sum_j \phi_j\right)\log\theta + \left(d - \sum_j \phi_j\right)\log(1-\theta)\right)$$
$$= \exp\left(\left(\alpha_0 - 1 + \sum_j \phi_j\right)\log\theta + \left(\beta_0 + d - \sum_j \phi_j - 1\right)\log(1-\theta)\right).$$

This implies $q^\star(\theta)$ follows a beta distribution with parameters:

$$\alpha = \alpha_0 + \sum_j \phi_j,$$

$$\beta = \beta_0 + d - \sum_j \phi_j.$$

Update for v and V: Based on Eqn. 4.6, the exponentiated conditional expectation of all the parameters except $\mathbf{\Sigma}_0^{-1}$ up to a constant scaling factor:

$$q^\star\left(\mathbf{\Sigma}_0^{-1}\right) \propto \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{V}_0^{-1}\mathbf{\Sigma}_0^{-1}\right)\right) - \frac{v_0 + d - T - 1}{2}\log\det\left(\mathbf{\Sigma}_0^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}_0^{-1}\left(\sum_j \boldsymbol{m}_{(j)}\boldsymbol{m}_{(j)}^\top + \mathbf{\Sigma}_j\right)\right).$$

This implies $q^\star\left(\mathbf{\Sigma}_0^{-1}\right)$ follow a Wishart distribution with parameters:

$$v = v_0 + d,$$

$$\boldsymbol{V} = \left(\boldsymbol{V}_0^{-1} + \sum_j \boldsymbol{m}_{(j)}\boldsymbol{m}_{(j)}^\top + \mathbf{\Sigma}_j\right)^{-1}.$$

Update for $\mathbf{\Sigma}_j$: all the terms involve $\mathbf{\Sigma}_j$ in ELBO approximation (Eqn. 4.5):

$$-\frac{1}{2}\operatorname{tr}\left(v\boldsymbol{V}\mathbf{\Sigma}_j\right) - \frac{1}{8}\sum_t \left(\mathbf{\Sigma}_j\right)_{t,t}\phi_j \sum_i \left(x_t^{ij}\right)^2 + \frac{1}{2}\log\det\left(\mathbf{\Sigma}_j\right).$$

Rewrite the second term:

$$-\frac{1}{8}\operatorname{tr}\left(\mathbf{\Sigma}_j \operatorname{diag}\left(\left[\textstyle\sum_i \phi_j\left(x_1^{ij}\right)^2, \cdots, \sum_i \phi_j\left(x_T^{ij}\right)^2\right]^\top\right)\right).$$

Denote the diagonal matrix as $\tilde{\boldsymbol{X}}_j$. For every j, we have a constrained optimization problem:

$$\max_{\mathbf{\Sigma} \in \mathbb{S}_{++}^T} \log\det\left(\mathbf{\Sigma}\right) - \operatorname{tr}\left(\mathbf{\Sigma}\left(v\boldsymbol{V} + \frac{1}{4}\tilde{\boldsymbol{X}}_j\right)\right).$$

which admits a closed form solution:

$$\mathbf{\Sigma}_j^\star = \left(v\boldsymbol{V} + \frac{1}{4}\tilde{\boldsymbol{X}}_j\right)^{-1}. \tag{4.7}$$

for $j = 1, \cdots d$.

Update for $m_{(j)}$: all the terms involved $m_{(j)}$ in ELBO approximation (Eqn. 4.5):

$$
\begin{aligned}
f\left(\boldsymbol{m}_{(j)}\right) := & -\frac{v}{2}\left\langle \boldsymbol{m}_{(j)}, \boldsymbol{V}\boldsymbol{m}_{(j)}\right\rangle - \frac{1}{8}\left\langle \boldsymbol{m}_{(j)}, \tilde{\boldsymbol{X}}_j \boldsymbol{m}_{(j)}\right\rangle \\
& + \left\langle \boldsymbol{m}_{(j)}, \left[\phi_j \textstyle\sum_i \left(y_1^i - \tilde{y_1^i}\right) x_1^{ij}, \quad \cdots, \quad \phi_j \sum_i \left(y_T^i - \tilde{y_T^i}\right) x_T^{ij}\right]^\top \right\rangle \\
& + \frac{1}{4}\left\langle \boldsymbol{m}_{(j)}, \left[\phi_j \textstyle\sum_i \left\langle \boldsymbol{w}'_1 \circ \boldsymbol{z}', \boldsymbol{x}_1^i\right\rangle x_1^{ij}, \quad \cdots, \quad \sum_i \left\langle \boldsymbol{w}'_T \circ \boldsymbol{z}', \boldsymbol{x}_T^i\right\rangle x_T^{ij}\right]^\top \right\rangle \\
& - \frac{1}{4}\left\langle \boldsymbol{m}_{(j)}, \left[\phi_j \textstyle\sum_i x_1^{ij} \sum_{l\neq j} \phi_l x_1^{il} m_{1l}, \quad \cdots, \quad \phi_j \sum_i x_T^{ij} \sum_{l\neq j} \phi_l x_T^{il} m_{tl}\right]^\top \right\rangle .
\end{aligned}
$$

This problem is quadratic with a negative definite Hessian matrix, hence by stationary condition (i.e zero gradient) we have closed form updates:

$$
\begin{aligned}
\boldsymbol{m}_{(j)} = \boldsymbol{\Sigma}_j^\star (& \left[\phi_j \textstyle\sum_i \left(y_1^i - \tilde{y_1^i}\right) x_1^{ij}, \quad \cdots, \quad \phi_j \sum_i \left(y_T^i - \tilde{y_T^i}\right) x_T^{ij}\right]^\top \\
& + \frac{1}{4}\left[\phi_j \textstyle\sum_i \left\langle \boldsymbol{w}'_1 \circ \boldsymbol{z}', \boldsymbol{x}_1^i\right\rangle x_1^{ij}, \quad \cdots, \quad \sum_i \left\langle \boldsymbol{w}'_T \circ \boldsymbol{z}', \boldsymbol{x}_T^i\right\rangle x_T^{ij}\right]^\top \\
& - \frac{1}{4}\left[\phi_j \textstyle\sum_i x_1^{ij} \sum_{l\neq j} \phi_l x_1^{il} m_{1l}, \quad \cdots, \quad \phi_j \sum_i x_T^{ij} \sum_{l\neq j} \phi_l x_T^{il} m_{tl}\right]^\top).
\end{aligned} \tag{4.8}
$$

For all $j = 1 \cdots d$. When the reference point of quadratic lower bound $\boldsymbol{w}' \circ \boldsymbol{z}'$ is set to be the mean parameters from the previous iteration, we can simplify Eqn. 4.9:

$$
\begin{aligned}
\boldsymbol{m}_{(j)}^{(k+1)} = \boldsymbol{\Sigma}_j^\star (& \left[\phi_j \textstyle\sum_i \left(y_1^i - \tilde{y_1^i}\right) x_1^{ij}, \quad \cdots, \quad \phi_j \sum_i \left(y_T^i - \tilde{y_T^i}\right) x_T^{ij}\right]^\top \\
& + \frac{1}{4}\left[\phi_j^2 \textstyle\sum_i \left(x_1^{ij}\right)^2 m_{1j}^{(k)}, \quad \cdots, \quad \sum_i \phi_j^2 \sum_i \left(x_T^{ij}\right)^2 m_{Tj}^{(k)}\right]^\top .
\end{aligned} \tag{4.9}
$$

Update for ϕ_j : All the terms involve ϕ_j in ELBO approximation (Eqn. 4.5):

$$\begin{aligned} f\left(\phi_j\right) := & \phi_j\left(\psi\left(\alpha\right)-\psi\left(\beta\right)\right)+\phi_j \sum_t \sum_i \left(y_t^i-\tilde{y_t^i}\right) m_{tj} x_t^{ij} \\ & +\frac{\phi_j}{4} \sum_t \sum_i \left(m_{tj} x_t^{ij}\right)\left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle \\ & -\frac{\phi_j}{4} \sum_t \sum_i m_{tj} x_t^{ij} \sum_{l \neq j} m_{tl} \phi_l x_t^{il}-\frac{\phi_j}{8} \sum_t \left(\left(\Sigma_j\right)_{tt}+m_{tj}^2\right) \sum_i \left(x_t^{ij}\right)^2 \\ & -\phi_j \log\left(\phi_j\right)-\left(1-\phi_j\right) \log\left(1-\phi_j\right). \end{aligned}$$

Observe $f\left(\phi_j\right)$ is a smooth strictly concave function, so we can solve for $\phi_j^\star$ by stationary condition (i.e 0 derivatives), which admit a closed form update:

$$\phi_j^\star = \sigma\left(a\right). \tag{4.10}$$

where $\sigma\left(t\right) := \frac{1}{\exp(-t)+1}$ denote the sigmoid function, and:

$$\begin{aligned} a = & \psi\left(\alpha\right)-\psi\left(\beta\right)+\sum_t \sum_i \left(y_t^i-\tilde{y_t^i}\right) m_{tj} x_t^{ij} \\ & +\frac{1}{4} \sum_t \sum_i \left(m_{tj} x_t^{ij}\right)\left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle \\ & -\frac{1}{4} \sum_t \sum_i m_{tj} x_t^{ij} \sum_{l \neq j} m_{tl} \phi_l x_t^{il}-\frac{1}{8} \sum_t \left(\left(\Sigma_j\right)_{tt}+m_{tj}^2\right) \sum_i \left(x_t^{ij}\right)^2. \end{aligned}$$

When reference point of quadratic lower bound $\boldsymbol{w}' \circ \boldsymbol{z}'$ is set to be the mean parameters from the previous iterations, Eqn. 4.10 is simplified:

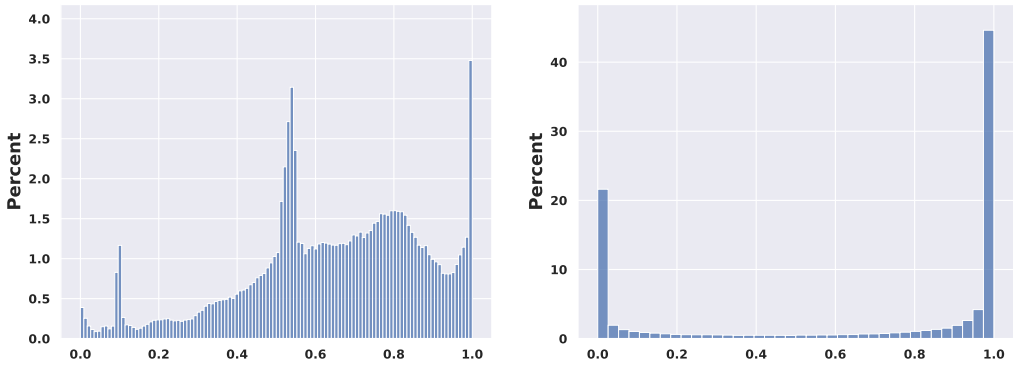
$$\phi_j^{k+1} = \sigma\left(a\right). \tag{4.11}$$

where

$$\begin{aligned} a = & \psi\left(\alpha\right)-\psi\left(\beta\right)+\sum_t \sum_i \left(y_t^i-\tilde{y_t^i}\right) m_{tj} x_t^{ij} \\ & +\frac{1}{8} \sum_t \left(m_{tj}^2\left(2\phi_j^{(k)}-1\right)-\left(\Sigma_j\right)_{tt}\right) \sum_i \left(x_t^{ij}\right)^2. \end{aligned}$$

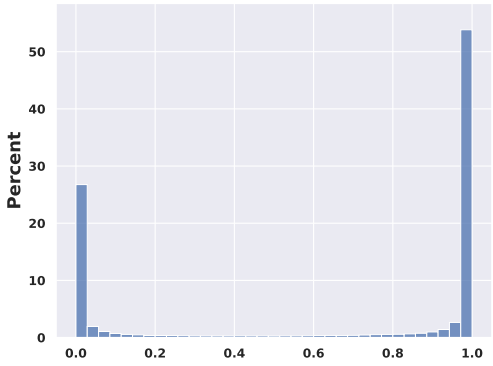
4.6.2 Additional Experimental Results

This subsection include the additional experimental results on the microbiome data from Section 4.4.2: Fig. 4.3 and Fig. 4.4 include the predicted probabilities on training and test data for the other taxon levels respectively, Fig. 4.5 includes the additional calibration curves, and Fig. 4.6, Fig. 4.7, Fig. 4.8, Fig. 4.9, Fig. 4.10 and Fig. 4.11 include feature sparsity plots for Kingdom, Phylum, Class, Genus and Species respectively.

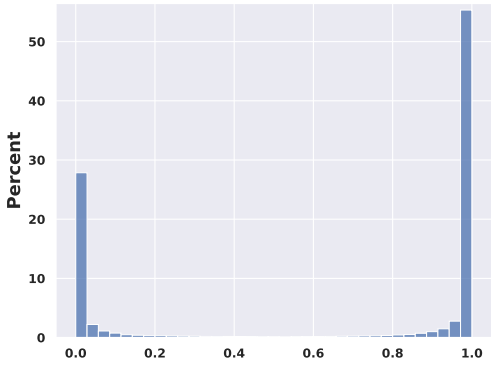


(a) Kingdom

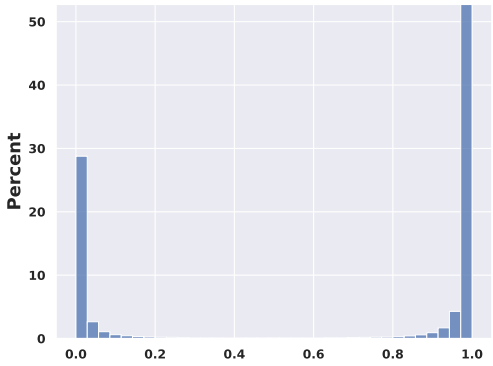
(b) Phylum



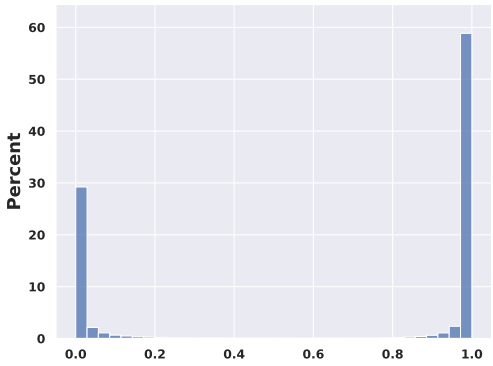
(c) Class



(d) Family

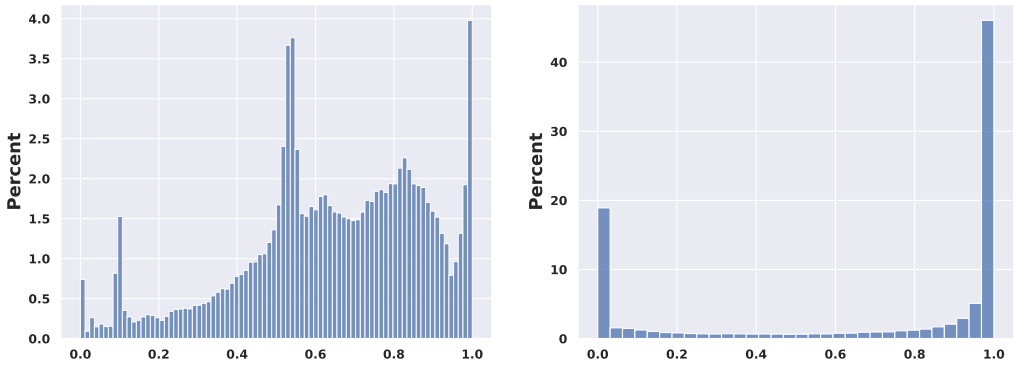


(e) Genus



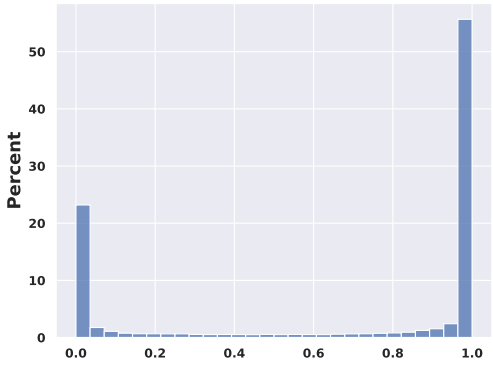
(f) Species

Figure 4.3: Histogram of predicted probabilities on training data for different Taxon levels.

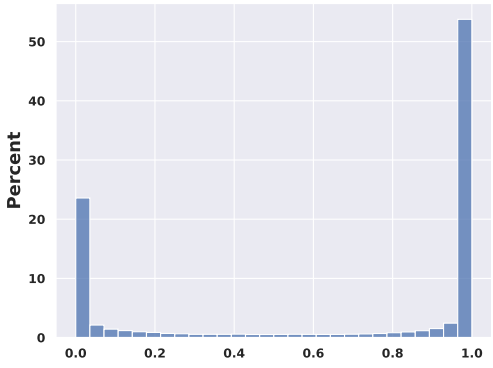


(a) Kingdom

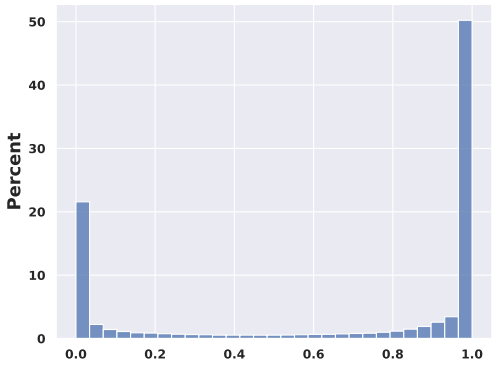
(b) Phylum



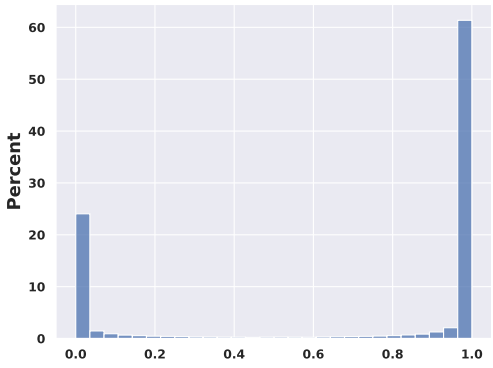
(c) Class



(d) Family

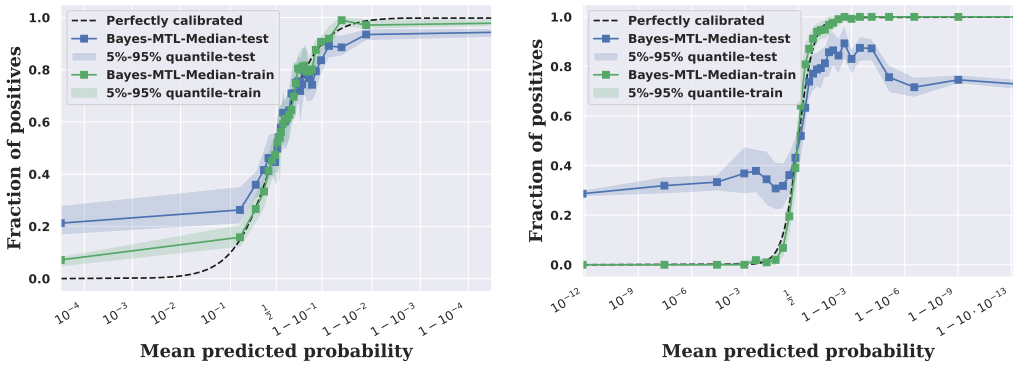


(e) Genus



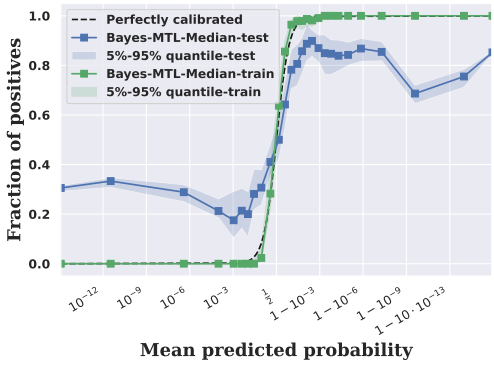
(f) Species

Figure 4.4: Histogram of predicted probabilities on test data for different Taxon levels.

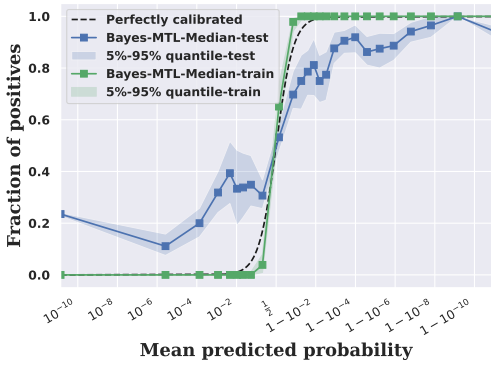


(a) Kingdom

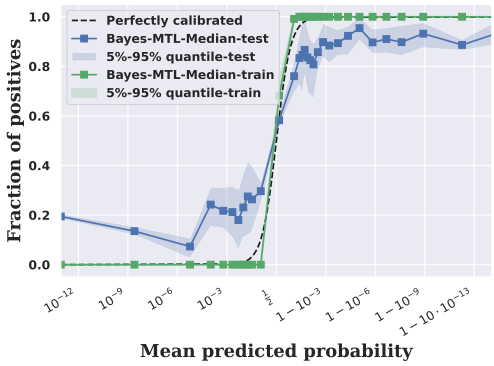
(b) Phylum



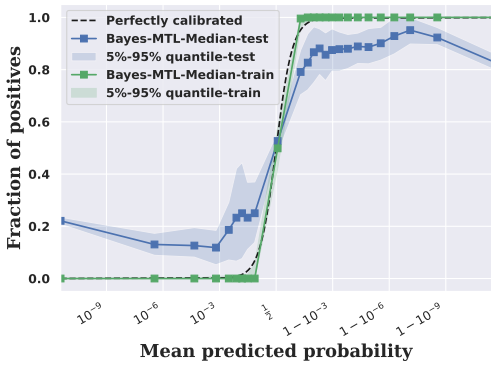
(c) Class



(d) Family

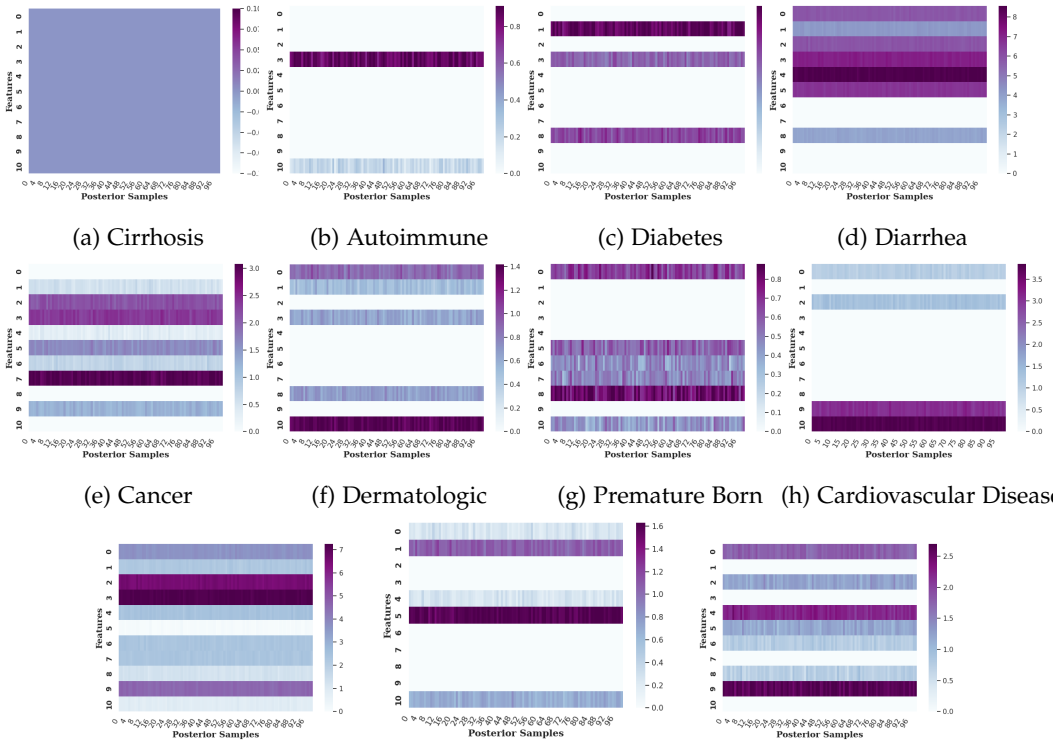


(e) Genus



(f) Species

Figure 4.5: Calibration curves for different Taxon levels.

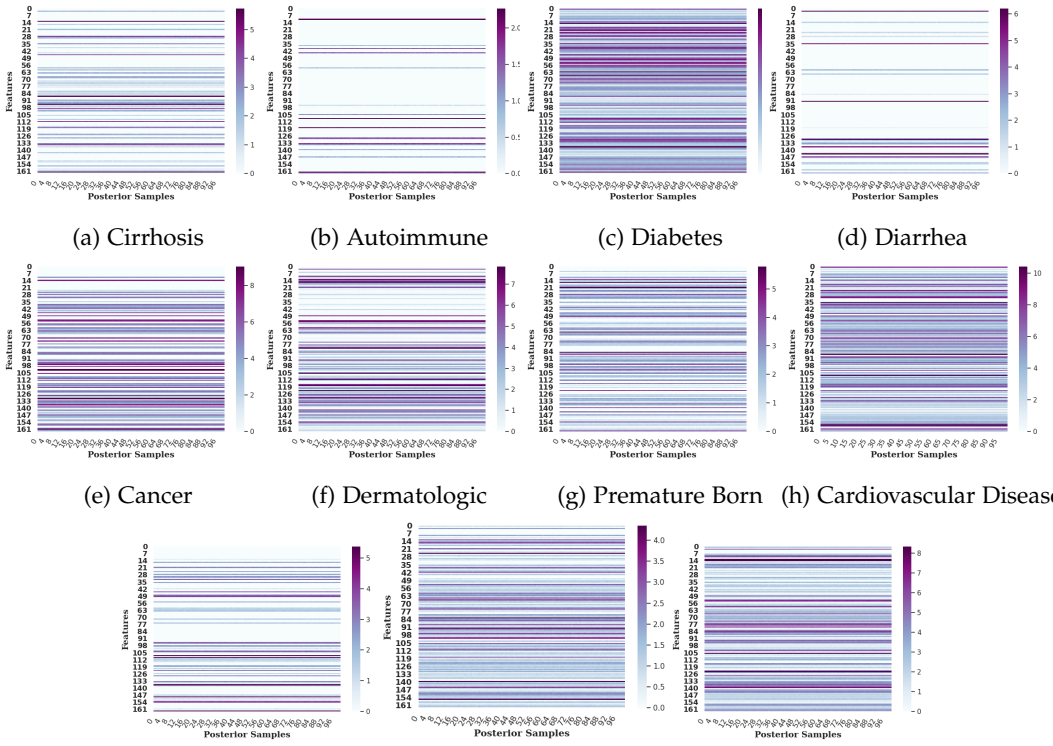


(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.6: Feature importance weight visualization across 11 different disease category of Kingdom taxon level.

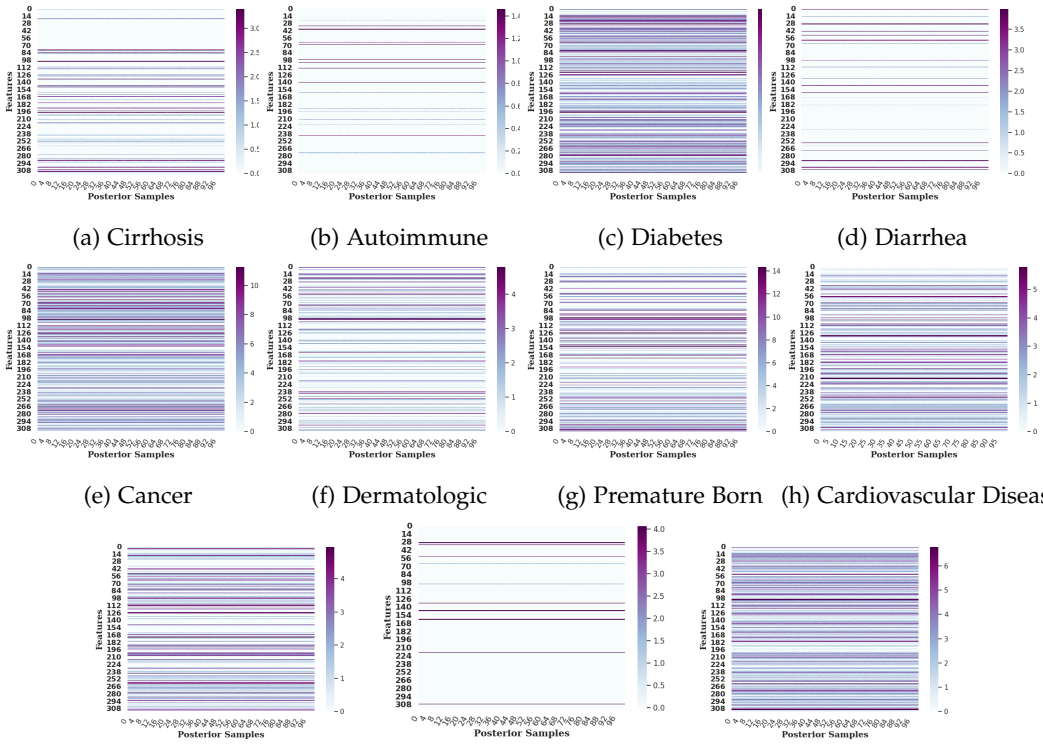


(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.7: Feature importance weight visualization across 11 different disease category of Phylum taxon level.

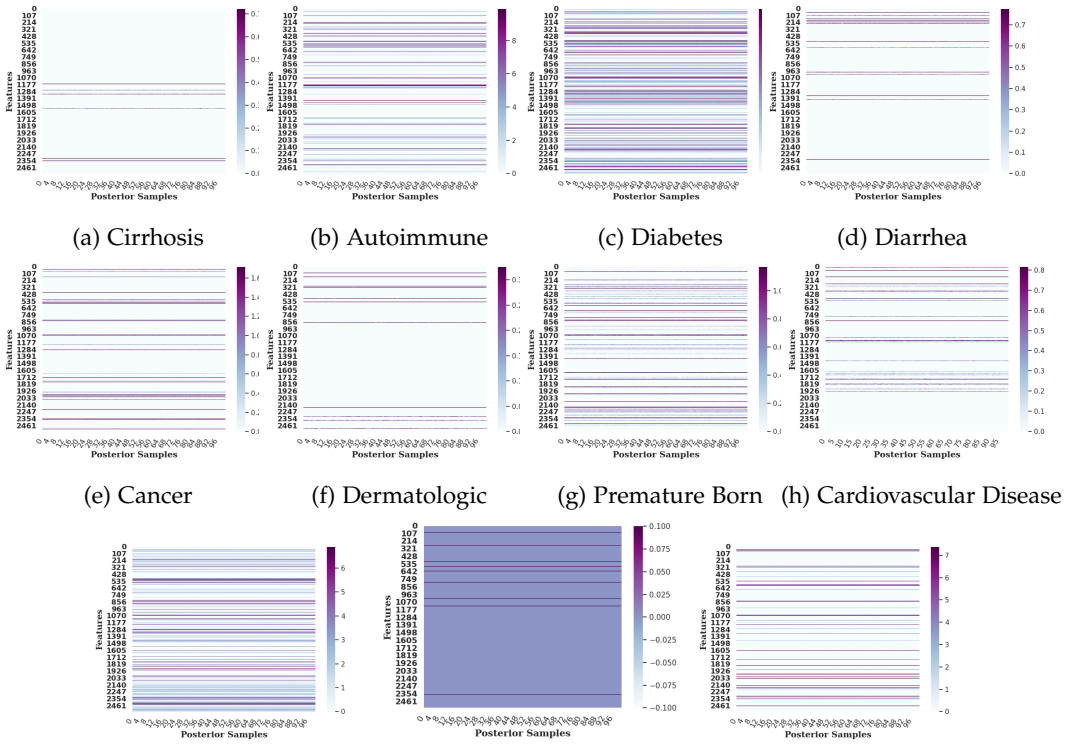


(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.8: Feature importance weight visualization across 11 different disease category of Class taxon level.

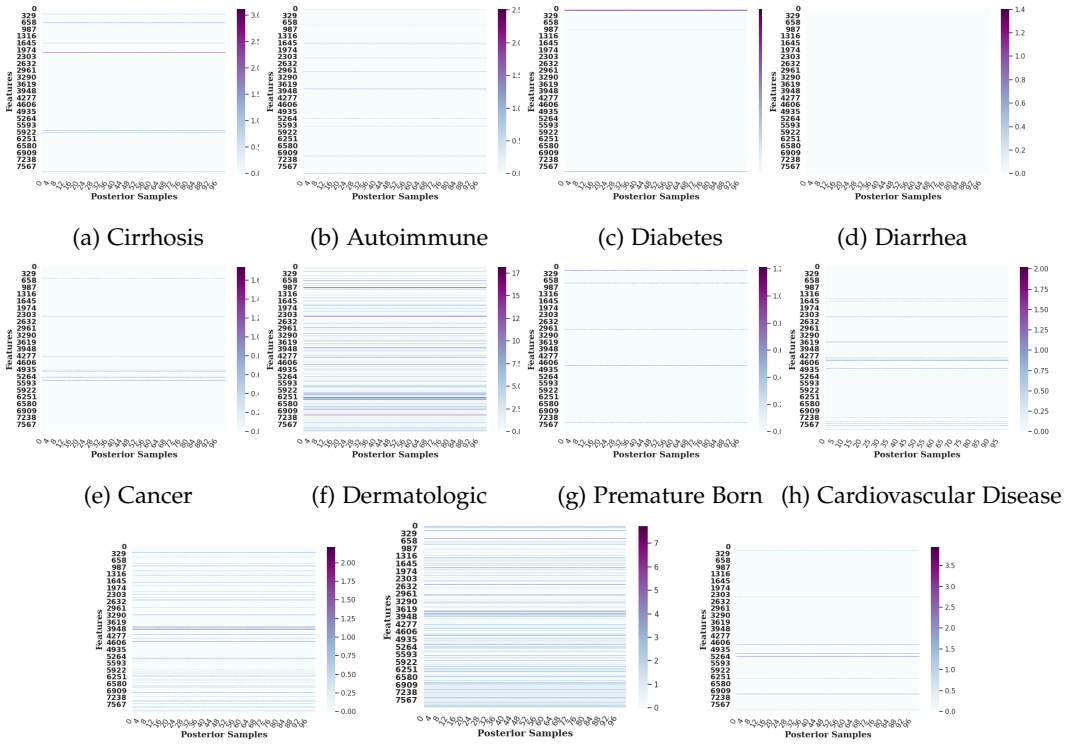


(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.9: Feature importance weight visualization across 11 different disease category of Family taxon level.

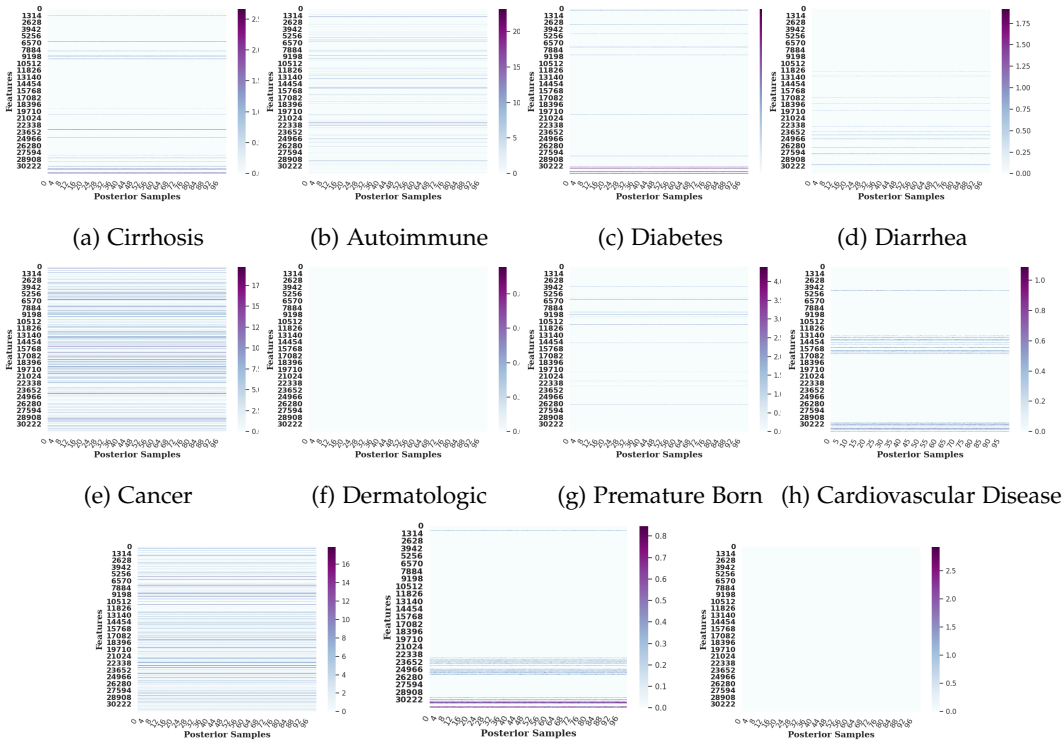


(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.10: Feature importance weight visualization across 11 different disease category of Genus taxon level.



(a) Cirrhosis (b) Autoimmune (c) Diabetes (d) Diarrhea

(e) Cancer (f) Dermatologic (g) Premature Born (h) Cardiovascular Disease

(i) Neurological Disease (j) Gastrointestinal Infection (k) Bowel Disease

Figure 4.11: Feature importance weight visualization across 11 different disease category of Species taxon level.

CHAPTER V

Recovery of Transition Probabilities from Marginals of Two-Way Tabular Data

5.1 Introduction

In this chapter, we formulate the transition matrix recovery problem as estimation of a stochastic matrix of conditional probabilities from multiple experiments, where we are given multiple two-way contingency tables with known margin sums but missing inner cells. This formulation has application to a wide range applications including bird migration [189], recommender systems [190], voting data analysis [191, 192] which will be further discussed below in addition to applications in credit risk modeling [193].

5.1.1 Related Work and Applications

Our formulation is related to estimation of the state transition matrix of a time homogeneous discrete state Markov process when the data is temporal [194]. For this problem the method of conditional least squares (CLS) [195, 196, 197, 198, 199] is commonly applied to estimate the state transition matrix and its performance has been analyzed both theoretically and in simulation [194]. However, CLS is poorly matched to maximum likelihood estimator unless the data is Gaussian distributed which is not the case in categorical setting we are interested in. Our

formulation can also be related to probabilistic graphical models, in particular the collective graphical model [189, 200] where individual data are assumed to be sampled from a graphical model. However such models make additional assumptions on the underlying (graphical) dependency structure between the features (categories) of the data. In contrast, our proposed approach makes no such assumptions on dependency structure. Next we describe two specific application domains of our work and its relation to contingency table analysis.

Recommender systems. The transition matrix recovery problem studied here has application to recommender systems in e-commerce, in particular to sequence-aware recommender systems [190] and next-basket-analysis [201]. A recommender system collects a matrix of counts of clicks on links, i.e., click through rates, on web pages collected from a number of users, to build a profile of all the users. This profile is used to target advertising to a particular user based on collaborative filtering. In sequence aware recommender systems pairs of successive clicks, which could be on the same or different webpages, are recorded and the estimated transition matrix is used to refine the user profile to improve the pitch of targeted advertising. Often, in e-commerce users are anonymous, which prevents tracking of individual ID's and thus only marginal data is available. Thus the transition matrix is not directly observable and must be estimated from the marginals. Thus the results in this chapter are directly applicable to such recommender systems.

Political election voting analysis. Our formulated transition matrix estimation problem has application to analysis of voting systems [202, 203, 204]. For example, in political science, exit polls often collect information on how voters voted in an

election. They may also ask the voter how she voted in the last election or what party affiliations were held by the candidates she voted for in a multi-category election, e.g., an election for state legislature, federal congress, and presidential candidates. Exit poll voting data is often separately aggregated according to electoral districts and individual level voter-specific cross-tabulated data may not be reported. In this case only marginal data per district is available and the results in this chapter can be applied to recover the conditional probability (transition) matrices associated with voter choices across the election categories. This application will be illustrated with simulated and real data experiments in Section 5.4.3 of this chapter.

Contingency table analysis. The transition matrix recovery problem we address can be related to contingency table analysis. A two way contingency table is a matrix of tabular data where each matrix entry, called a cell, corresponds to the number of counts that occur in a pair of outcome categories. In classical contingency table analysis the principal objective is to test hypotheses on the probabilistic relations, called the "cell probabilities," between the row and column covariates given empirical cell count data [205, 206]. In contrast, the objective in this work is to recover the stochastic matrix (conditional probabilities of row counts given observed count levels of the column variables) from marginal data collected from multiple independent and related contingency tables. In the sequel we will formulate the transition matrix recovery problem using the terminology of contingency tables.

5.1.2 Contributions and Organization

The main contributions of this chapter includes: an exact model with minimal assumptions for the transition matrix recovery problem, three valid approximations of the exact model and a novel Riemannian gradient algorithm with Polyak adaptive step size to obtain the Maximum Likelihood Estimatiors (MLE) of the transition matrix.

The proposed methods are applied to a synthetic dataset and a real world dataset from New Zealand general election [207] in comparison with CLS [195]. Our results show the scopes when those approximation apply. A further clustering analysis using the estimated stochastic matrices across different electorate districts is able to identify communities that are reflective of the demographics of New Zealand.

The remainder of this chapter is organized as following: Section 5.2 formally introduces the mathematical formulation of the transition matrix recovery problem, the exact model along with three approximations in connection to CLS [195], Section 5.3 introduces the proposed Riemannian gradient algorithm to obtain the Maximum Likelihood Estimators (MLE), Section 5.4 includes application of the methods to a synthetic dataset and the New Zealand general election dataset [207], and Section 5.5 summarizes the major takeaways from the chapter.

5.2 Proposed Model

5.2.1 Mathematical formulation

Notations: We use bold upper case letters for matrices, bold lower case letters for vectors and no bold lower case for scalars. The Hadamard (element-wise) product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $\boldsymbol{a} \circ \boldsymbol{b}$, and Hadamard (element-wise)

division is denoted by $\boldsymbol{a} \oslash \boldsymbol{b}$. diag denotes the function map a vector to a diagonal matrix with the vector as its diagonal entries, and diagnal denotes the function map a square matrix to a vector of its diagonal entries. We denote Δ_m for the m-dimensional probability simplex (i.e non-negative vectors of dimensional m that sum to 1) and $\Delta_{m\times d}$ for the space of stochastic matrices of dimension $m \times d$ with columns in m-dimensional probability simplex. $\boldsymbol{I}_m$ denotes the identity matrix of dimension m, and $\mathbf{1}_m$ denotes the all-ones vector in dimension m.

Problem Description: Data is collected from n independent experiments on choices of N_i individuals, $i = 1, \ldots, n$. In each experiment, an individual selects an item k from Category 1, and an item j from Category 2, where $k \in \{1, \ldots, d\}$ and $j \in \{1, \ldots, m\}$. The experimenter only observe the marginal count data for each experiment, i.e the histogram of selection counts from Category 1 and Category 2, respectively. More precisely, for all the experiments ranging from $i = 1, \cdots n$, we denote the observed marginal count data as $\boldsymbol{x}_i \in \mathbb{N}^d$ for category 1 and $\boldsymbol{y}_i \in \mathbb{N}^m$ for category 2 . In matrix form the marginal count data can be expressed as two matrices $\boldsymbol{X} := \begin{bmatrix} | & | & | \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \\ | & | & | \end{bmatrix} \in \mathbb{N}^{d\times n}$ and $\boldsymbol{Y} := \begin{bmatrix} | & | & | \\ \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_n \\ | & | & | \end{bmatrix} \in \mathbb{N}^{m\times n}$. The total population count, i.e the number of individuals participating in the i-th experiment or equivalently the sum of i-th column of matrix $\boldsymbol{X}$ or $\boldsymbol{Y}$, which are identical, is N_i. Of interest to the experimenter is the conditional probability that an individual selects item j in Category 2 given that they select item k in Category 1. The conditional probabilities can be estimated from the individual cross tabular data. However, the experimenter only observes population level marginal data.

Stated more concretely, the experimenter is interested in estimating the stochastic matrix $\mathbf{\Pi} \in \Delta_{m \times d}$ of conditional probabilities $\Pi_{jk} = P(Y = j \mid X = k)$ where $j \in \{1, \cdots, m\}$ and $k \in \{1, \cdots, d\}$. We denote each columns of $\mathbf{\Pi}$ as $\boldsymbol{\pi}_j$ for $j = 1, \cdots, d$. See fig. 5.1 for a visualization of the observed data.

Figure 5.1 Example of available data to the transition matrix recovery problem. "?" means the cell value is missing. The columns of contingency tables corresponds to the dinstinct items of Cateogory 1 and rows corresponds to dinstinct items of Category 2.

Category 2 \ Category 1	Item 1	...	Item d	Total
Item 1	?	?	?	1500
⋮	⋮	⋮	⋮	⋮
Item m	?	?	?	2000
Total	1000	⋯	500	6000

⋮

Category 2 \ Category 1	Item 1	...	Item d	Total
Item 1	?	?	?	1000
⋮	⋮	⋮	⋮	⋮
Item m	?	?	?	4000
Total	800	⋯	1300	10000

5.2.2 The Exact Model

We make the following homogeneity assumption: given that for an individual that selects k-th item of Category 1, her selection in Category 2 is identical and independently distributed as a categorical distribution with probability vector $\boldsymbol{\pi}_k$ which does not depend on the experiment index $i = 1, \cdots, n$. Hence conditioned on Category 1 item counts $\boldsymbol{x}_i$, the Category 2 item counts $\boldsymbol{y}_i$ are given by:

$$\begin{aligned} \boldsymbol{z}_k \mid x_i &\sim \mathrm{Mul}\left(x_{ki}, \boldsymbol{\pi}_k\right) \quad \forall k = 1, \cdots, d. \\ \boldsymbol{y}_i \mid \boldsymbol{x}_i &= \sum_{k=1}^{d} \boldsymbol{z}_k. \end{aligned} \tag{5.1}$$

where x_{ki} denotes the counts over items in Category 1 observed in the i-th experiment, $k = 1, \ldots, d$, $i = 1, \cdots, n$, $\mathrm{Mul}\left(N, \boldsymbol{p}\right)$ denotes the multinomial distribution with total count N and probability vector $\boldsymbol{p}$, and the hidden variables z_k represent the counts over items in Category 2 from all the individuals who selected item k in Category 1. The resulting compound distribution is a special case of Poisson multinomial distribution (generalized multinomial distribution) [208, 209, 210]. Since the evaluation of this distribution function has combinatorical complexity [210], inference with respect to this exact model is computationally intractable, which is the main motivation for approximations introduced n the following subsection.

5.2.3 Likelihood Approximations

In this subsection, we present 4 different ways to approximate the intractable negative log-likelihood function of the model proposed in Eqn. 5.1 including the classical conditional least squares (CLS) [195] and the conditions when those approximations are applicable.

Conditional Least Squares

Conditional least squares (CLS) is based on the observation that, conditioned on $\boldsymbol{x}_i$ the expected value of $\boldsymbol{y}_i$ is $\boldsymbol{\Pi}\boldsymbol{x}_i$. Thus, a method of moments estimator is naturally the solution of the constrained least squares problem:

$$\min_{\boldsymbol{\Pi} \in \Delta_{m \times d}} f\left(\boldsymbol{\Pi}\right). \tag{5.2}$$

where $f(\mathbf{\Pi}) = \|\mathbf{\Pi}\boldsymbol{X} - \boldsymbol{Y}\|_F^2$. This approximation is valid when population size N_i is sufficient large and the covariance matrix is nearly homogeneous (i.e a constant factor times identity matrix).

Normal Approximation

As a consequence of central limit theorem [209, 210], when the population size (N_i) is sufficient large, the likelihood function of the model in Eqn. 5.1 can be well approximated by a degenerate multivariate Gaussian distribution with following mean and covariance matrix:

$$\begin{aligned} \boldsymbol{\mu}_i &= \mathbf{\Pi}\boldsymbol{x}_i, \\ \mathbf{\Sigma}_i &= \operatorname{diagnal}(\mathbf{\Pi}\boldsymbol{x}_i) - \mathbf{\Pi}\operatorname{diagnal}(\boldsymbol{x}_i)\mathbf{\Pi}^\top. \end{aligned} \tag{5.3}$$

This leads to negative conditional log-likelihood function, up to an unimportant constant:

$$l(\boldsymbol{Y}|\boldsymbol{X};\mathbf{\Pi}) = \sum_{i=1}^{n}\left(\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^\dagger (\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \frac{1}{2}\operatorname{plog\,det}(\mathbf{\Sigma}_i)\right). \tag{5.4}$$

where $\mathbf{\Sigma}^\dagger$ denotes the pseudo inverse and plog det denotes the pseudo log-determinant.

As discussed in [210], an alternative formulation to work around the degeneracy is to discard the last element of each $\boldsymbol{y}_i$ since conditioned on $\boldsymbol{x}_i$ the summation of $\boldsymbol{y}_i$ is fixed and the last entry is determined by the fixed sum constraint. Denote the modified data as $\tilde{\boldsymbol{Y}}$ and the modified stochastic matrix $\tilde{\mathbf{\Pi}} := \begin{bmatrix} \boldsymbol{I}_{m-1} \\ \mathbf{0} \end{bmatrix} \mathbf{\Pi} \in \mathbb{R}^{(m-1)\times d}$ ($\mathbf{\Pi}$ with last row removed). Then this leads to the non-degenerate

normal approximation with mean and covariance matrix:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_i &= \tilde{\boldsymbol{\Pi}}\boldsymbol{x}_i, \\ \tilde{\boldsymbol{\Sigma}}_i &= \text{diagnal}\left(\tilde{\boldsymbol{\Pi}}\boldsymbol{x}_i\right) - \tilde{\boldsymbol{\Pi}}\,\text{diagnal}\left(\boldsymbol{x}_i\right)\tilde{\boldsymbol{\Pi}}^\top. \end{aligned} \tag{5.5}$$

resulting in the negative conditional log-likelihood function, up to a constant:

$$l\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right) = \sum_{i=1}^{n}\left(\frac{1}{2}\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{\mu}}_i\right)^\top \tilde{\boldsymbol{\Sigma}}_i{}^{-1}\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{\mu}}_i\right) + \frac{1}{2}\log\det\left(\tilde{\boldsymbol{\Sigma}}_i\right)\right). \tag{5.6}$$

Though Eqn. 5.4 and Eqn. 5.6 are well defined objective functions in theory, their evaluations run into numerical issues when $\boldsymbol{\Pi}$ is permutation similar a block-wise diagonal matrix due to rank deficiency in the covariance matrix. Instead, we propose to minimize:

$$f\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right) = l\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right) + \lambda\sum_{j=1}^{m}\sum_{k=1}^{d}\Pi_{jk}\log\left(\Pi_{jk}\right) \tag{5.7}$$

where $l\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right)$ is from either Eqn. 5.4 or Eqn. 5.6, and λ is a hyperparameter that controls the strength of regularization. This is known as entropy regularization and widely used in statistics and machine learning when working with discrete distributions [211, 212, 213].

Poisson Approximation

As a consequence of Poisson approximation theorems [214, 215, 208], when the entries of $\boldsymbol{x}_i$ are sufficiently large and the columns of stochastic matrix $\boldsymbol{\Pi}$ are close to uniform (i.e all entries have same values $\frac{1}{m}$), we can approximate the conditional distribution $\boldsymbol{y}_i \mid \boldsymbol{x}_i$ by a multivariate Poisson distribution with mean parameter $\boldsymbol{\Pi}\boldsymbol{x}_i$. This leads to a negative conditional log-likelihood function, up to a constant:

$$l\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} -y_{ji}\log\left(\sum_{k=1}^{d}\boldsymbol{\Pi}_{jk}\boldsymbol{x}_{ki}\right). \tag{5.8}$$

Multinomial Approximation

Alternatively we can approximate the Poisson multinomial distribution by a standard multinomial distribution, and the approximation will be exact when all the columns of the stochastic matrix are identical (i.e the two categories are independent). This leads to a negative conditional log-likelihood function, up to a constant:

$$l\left(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Pi}\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}-y_{ji}\log\left(\sum_{k=1}^{d}\boldsymbol{\Pi}_{jk}\frac{\boldsymbol{x}_{ki}}{\sum_{l=1}^{d}x_{li}}\right). \tag{5.9}$$

Observe that this is a weighted version of Eqn. 5.8.

5.3 Approximate Maximum Likelihood with Riemannian Gradient Algorithm

Due to the non-trivial composition structure of the negative likelihood functions in Eqns. (5.4), (5.6), (5.8), (5.9) together with the constraints that the columns of the transition matrix $\boldsymbol{\Pi}$ belong to the probability simplex, analytical expressions for the maximum likelihood solutions are intractable. Thus, we propose to use iterative methods to solve for the maximum likelihood estimators. In particular, we use a novel Riemannian gradient algorithm with adaptive step size to exploit the geometric structure of the underlying simplex constraint (Algorithm. 4). The adaptive step size we propose is an extension of previous work [216, 217, 218] to the Riemannian manifold based on the Polyak step size rule [216]. Section 5.3.1 introduces the general purpose algorithm and Section 5.3.2 summarize the gradient computations of the negative log-likelihood functions introduced in Section 5.2.3.

5.3.1 Riemannian Gradient Algorithm

In this subsection, we present the novel Riemannian gradient algorithm with adaptive Polyak step size, and this is a general purpose algorithm applicable beyond the transition matrix recovery problem considered in this chapter.

The Riemannian gradient algorithm [219, 220] is a class of iterative methods that solve the constrained optimization problem:

$$\min_{\boldsymbol{u}\in\mathcal{M}} f(\boldsymbol{u}). \tag{5.10}$$

where $\mathcal{M}$ is Riemannian manifold and f is a smooth function. It is an extension of the standard gradient descent algorithm to the Riemannian manifold. In the Riemannian gradient algorithm, the standard euclidean gradient is replaced with Riemannian gradient (Def. V.1), and instead of performing an update at every iteration based on line segment ($\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta\nabla f(\boldsymbol{u}_t)$), Riemannian gradient algorithm moves along the parameterized curve on the manifold defined by the retraction map (Def. V.2). The formal definitions of Riemannian gradient and retraction map are given below:

Definition V.1 (Riemannian Gradient on an Embedded Manifold of $\mathbb{R}^d$, Chapter 3.8 of [220])**.** *Given a smooth function f on an embedded submanifold manifold $\mathcal{M} \subseteq \mathbb{R}^d$ with standard Euclidean metric, then the Riemannian gradient of f at location $\boldsymbol{u}$ is defined as:*

$$grad\, f(\boldsymbol{u}) := \mathrm{proj}_{T_{\boldsymbol{u}}(\mathcal{M})}\left(\boldsymbol{\nabla} f(\boldsymbol{u})\right). \tag{5.11}$$

where $\boldsymbol{\nabla} f(\boldsymbol{u})$ is the standard Euclidean gradient, and $\mathrm{proj}_{T_{\boldsymbol{u}}(\mathcal{M})}$ is the projection operator maps a vector to $T_{\boldsymbol{u}}(\mathcal{M})$ (tangent space at point $\boldsymbol{u}$).

Definition V.2 (Retraction Map, Chapter 3.6 of [220]). *A Retraction on a manifold $\mathcal{M}$ at location $\boldsymbol{u}$ is a smooth map:*

$$R_{\boldsymbol{u}} : T_{\boldsymbol{u}}\mathcal{M} \to \mathcal{M} : \boldsymbol{v} \to R_{\boldsymbol{u}}(\boldsymbol{v}) \tag{5.12}$$

such that each curve $c(t) = R_{\boldsymbol{u}}(t\boldsymbol{v})$ satisfies $c(0) = \boldsymbol{u}$ and $c'(0) = \boldsymbol{v}$. This is a first order approximation to Geodesic (i.e the shortest path between points on the manifold), and the parameterized curve $c(t)$ is the extension of line segment in the Euclidean case.

With proper choice of step size (η_t), the vanilla Riemannian gradient algorithm is:

$$\boldsymbol{u}_{t+1} = R_{\boldsymbol{u}_t}\left(-\eta_t grad\, f(\boldsymbol{u}_t)\right). \tag{5.13}$$

In the particular case of $\Delta_{m\times d}$ considered in this chapter, Riemannian Gradient of a smooth function f with Euclidean gradient function $\boldsymbol{\nabla} f$ is:

$$grad\, f(\boldsymbol{\Pi}) = \boldsymbol{\Pi} \circ \boldsymbol{\nabla} f(\boldsymbol{\Pi}) - \boldsymbol{\Pi}\,\text{diagnal}\left(\text{diag}\left(\boldsymbol{\nabla} f(\boldsymbol{\Pi})^\top \boldsymbol{\Pi}\right)\right). \tag{5.14}$$

For the negative likelihood functions in Eqs. (5.4), (5.6), (5.8), (5.9), the computations of Euclidean gradients are presented in the next subsection.

The retraction map $R_{\boldsymbol{\Pi}} : T_{\boldsymbol{\Pi}}(\Delta_{m\times d}) \to \Delta_{m\times d}$ at location $\boldsymbol{\Pi}$ is given by:

$$R_{\boldsymbol{\Pi}}(\boldsymbol{V}) := \boldsymbol{\Pi} \circ \exp(\boldsymbol{V})\,\text{diagnal}\left(\mathbf{1}_m^\top \boldsymbol{\Pi} \circ \exp(\boldsymbol{V})\right)^{-1}. \tag{5.15}$$

where exp is element-wise exponentiation, and this is a matrix extension of the exponentiated gradient or multiplicative weights [221].

The Polyak step size rule is first proposed in [216] in Euclidean setting, and is commonly used for subgradient method [217, 222]. Recently, [223, 218] extend it to stochastic optimization with both convergence guarantee and great empirical

success. In this section, we apply the Polyak step size rule to the Riemannian setting, where Algorithm 3 summarize the Riemannian gradient algorithm with Polyak step size with a fixed lower bound of the objective function, and Algorithm 4 is an extension of [217] to Riemannian setting where the lower bound estimate is iterative refined.

Algorithm 3 Riemannian Gradient algorithm with Polyak stepsize

Inputs: Data $(\boldsymbol{X}, \boldsymbol{Y})$, maximum number of iteration T,
objective function f (Eqn. 5.7, Eqn. 5.8, Eqn. 5.9),
objective lower bound $\tilde{f}$
Intermediate: step size η_t
Initialize: $\boldsymbol{\Pi}_0$
  for $t = 0, \cdots T-1$ **do**
    $\boldsymbol{g}_t \leftarrow grad\, f\left(\boldsymbol{\Pi}_t\right)$ (Eqn. 5.14)
    $\eta_t \leftarrow \min\left(\frac{f(\boldsymbol{\Pi}_t)-\tilde{f}}{2\|\boldsymbol{g}_t\|_2^2}, \frac{1}{\sqrt{t+1}}\right)$
    $\hat{\boldsymbol{\Pi}}_{t+1} \leftarrow R_{\boldsymbol{\Pi}_t}\left(-\eta_t \boldsymbol{\nabla}_t\right)$ (Eqn. 5.15)
  end for
  return MLE estimator $\boldsymbol{\Pi}_{t^\star}$, where $t^\star = \arg\min_{t<T}\{f\left(\boldsymbol{\Pi}_t\right)\}$

Algorithm 4 Riemannian Gradient algorithm with adaptive lower bound estimate

Inputs: Data $(\boldsymbol{X}, \boldsymbol{Y})$, initial objective lower bound $\tilde{f}_0$
maximum number of epoch K, maximum number of inner iterations T
Initialize: $\hat{\boldsymbol{\Pi}}$ randomly
  for $\tau = 0, \cdots, K-1$ **do**
    Let $\boldsymbol{\Pi}_{\tau+1}$ be the output of Algorithm. 3 using input $\boldsymbol{\Pi}_\tau, T, \tilde{f}_\tau$
    $\tilde{f}_{\tau+1} \leftarrow \frac{f(\boldsymbol{\Pi}_{\tau+1})+\tilde{f}_\tau}{2}$
  end for
  return MLE estimator $\boldsymbol{\Pi}_{\tau^\star}$, where $\tau^\star = \arg\min_{\tau<K}\{f\left(\boldsymbol{\Pi}_\tau\right)\}$

5.3.2 Gradient Computation

The Riemannian gradient in Eqn. 5.14 requires computation of Euclidean gradient, and the gradients of all the negative log-likelihood in Eqs. (5.4), (5.6), (5.8), (5.9) require matrix differential computations. The reader is referred to [224] for a detailed discussion on matrix differentiation rules.

Useful Differential Results: we first summarize a few useful differential results that will be used to compute the gradients. See [225, 226, 224] for derivations.

$$
\begin{aligned}
d\ \log\det\left(\boldsymbol{\Sigma}\right) &= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\right) d\,\boldsymbol{\Sigma}, \\
d\,\boldsymbol{\Sigma}^{-1} &= -\boldsymbol{\Sigma}^{-1}\left(d\,\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}, \\
d\ \operatorname{plog}\det\left(\boldsymbol{\Sigma}\right) &= \operatorname{tr}\left(\boldsymbol{\Sigma}^{\dagger}\right) d\,\boldsymbol{\Sigma}, \\
d\,\boldsymbol{\Sigma}^{\dagger} &= -\boldsymbol{\Sigma}^{\dagger}\left(d\,\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{\dagger} + \boldsymbol{\Sigma}^{\dagger}\boldsymbol{\Sigma}^{\dagger}\left(d\,\boldsymbol{\Sigma}\right)\left(\boldsymbol{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\dagger}\right) + \left(\boldsymbol{I} - \boldsymbol{\Sigma}^{\dagger}\boldsymbol{\Sigma}\right)\left(d\,\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{\dagger}\boldsymbol{\Sigma}^{\dagger}.
\end{aligned}
$$

For a smooth matrix-variate scalar function f, the following relates the matrix differential to the gradient:

$$
d\,f\left(\boldsymbol{A}\right) = \operatorname{tr}\left(\left(d\boldsymbol{A}\right)\boldsymbol{G}\right) \Longleftrightarrow \boldsymbol{\nabla} f\left(\boldsymbol{A}\right) = \boldsymbol{G}. \tag{5.16}
$$

Using the results above, we can compute following (Euclidean) gradients:

$$
\begin{aligned}
&\text{(For Eqn.5.4)} \quad \nabla f\left(\boldsymbol{\Pi}\right) = \sum_{i=1}^{n}\left(-\left(\boldsymbol{A}_i + \boldsymbol{\Sigma}_i^{\dagger}\right)\boldsymbol{\Pi}\,\operatorname{diagnal}\left(\boldsymbol{x}_i\right) + \left(0.5\boldsymbol{A}_i + 0.5\,\operatorname{diag}\left(\boldsymbol{\Sigma}_i^{\dagger}\right) + \boldsymbol{\Sigma}_i\left(\boldsymbol{\mu}_i - \boldsymbol{y}\right.\right.\right. \\
&\text{(For Eqn.5.6)} \quad \nabla f\left(\boldsymbol{\Pi}\right) = \sum_{i=1}^{n}\left(-\left(\tilde{\boldsymbol{A}}_i + \tilde{\boldsymbol{\Sigma}}_i{}^{-1}\right)\tilde{\boldsymbol{\Pi}}\,\operatorname{diagnal}\left(\boldsymbol{x}_i\right) + \left(0.5\tilde{\boldsymbol{A}}_i + 0.5\,\operatorname{diag}\left(\tilde{\boldsymbol{\Sigma}}_i{}^{-1}\right) + \tilde{\boldsymbol{\Sigma}}_i\left(\tilde{\boldsymbol{\mu}}\right.\right.\right. \\
&\text{(For Eqn.5.8)} \quad \nabla f\left(\boldsymbol{\Pi}\right) = -\left(\boldsymbol{Y} \oslash \left(\boldsymbol{\Pi}\boldsymbol{X}\right)\right)\boldsymbol{X}^{\top}, \\
&\text{(For Eqn.5.9)} \quad \nabla f\left(\boldsymbol{\Pi}\right) = -\left(\boldsymbol{Y} \oslash \left(\boldsymbol{\Pi}\boldsymbol{P}\right)\right)\boldsymbol{P}^{\top}.
\end{aligned}
$$

Where $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ are function of $\boldsymbol{\Pi}$ in Eqn. 5.3. Defining the quantity $\boldsymbol{R}_i = \left(\boldsymbol{y}_i - \boldsymbol{\mu}_i\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_i\right)^{\top}$, $\boldsymbol{A}_i$ is given by:

$$
\boldsymbol{A}_i = -\boldsymbol{\Sigma}_i^{\dagger}\boldsymbol{R}_i\boldsymbol{\Sigma}_i^{\dagger} + \left(\boldsymbol{I}_d - \boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i^{\dagger}\right)\boldsymbol{R}_i\boldsymbol{\Sigma}_i^{\dagger}\boldsymbol{\Sigma}_i^{\dagger} + \boldsymbol{\Sigma}_i^{\dagger}\boldsymbol{\Sigma}_i^{\dagger}\boldsymbol{R}_i\left(\boldsymbol{I}_d - \boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i^{\dagger}\right).
$$

$\tilde{\boldsymbol{\Pi}}$ and $\tilde{\boldsymbol{y}}_i$ are the modified representation discussed in Section 5.2.3, $\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i$ are

functions of $\tilde{\mathbf{\Pi}}$ defined in Eqn. 5.5, $\tilde{A}_i$ is given by:

$$\tilde{A}_i = -\tilde{\mathbf{\Sigma}}_i^{-1} \left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{\mu}}_i\right) \left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{\mu}}_i\right)^\top \tilde{\mathbf{\Sigma}}_i^{-1}.$$

$\boldsymbol{P}$ is the normalized version of data matrix $\boldsymbol{X}$ with entries $P_{ki} = \frac{x_{ki}}{\sum_{l=1}^{d} x_{li}}$.

5.4 Experiments

In this section, we evaluate the performance of our proposed likelihood based Riemannian methods on both the simulated data and the New Zealand general election data[207]. The conditional least squares (CLS), which is implemented by applying Algorithm 4 to Eqn. 5.2, is compared as as a benchmark.

5.4.1 Evaluation Metrics

Since the columns of stochastic matrices $\mathbf{\Pi}$ are in probability simplex, we evaluate the estimation performance by applying various metrics to each column and computing their averages. The list of metrics summarized in Table 5.3 includes Jensen Shannon Divergence, Hellinger Distance, mean square error, and two newly introduced metrics Maximum Index Rank Agreement (MIRA)[1] and Top Cumulative Probability Intersection (TCPI)[2].

[1]MIRA quantifies the agreement between mode of $\boldsymbol{p}$ and mode of $\hat{\boldsymbol{p}}$
[2]TCPI quantifies the amount overlap between the typical set of $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$

Table 5.3 Summary of Evaluation Measures in Terms of $\boldsymbol{p}, \hat{\boldsymbol{p}} \in \Delta_m$.

Evaluation Metric	Expression	Range
Jensen-Shannon Divergence (JSD)	$\mathrm{JSD}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{2}\sum_{j=1}^{m} p_j \log\left(\frac{2p_j}{p_j+\hat{p}_j}\right) + \frac{1}{2}\sum_{j=1}^{m} \hat{p}_j \log\left(\frac{2\hat{p}_j}{p_j+\hat{p}_j}\right)$	$[0, 1]$
Hellinger Distance (HD)	$\mathrm{HD}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{j=1}^{m}(\sqrt{p_j} - \sqrt{\hat{p}_j})^2}$	$[0, 1]$
Mean Square Error (MSE)	$\mathrm{MSE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{m}\sum_{j=1}^{m}(p_j - \hat{p}_j)^2$	$[0, \frac{2}{m}]$
Maximum Index Rank Agreement (MIRA)	$j^\star = \arg\max_j p_j$ $\mathrm{MIRA}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \left\lvert\{j \in \{1, \ldots, m\} : \hat{p}_j > \hat{p}_{j^\star}\}\right\rvert + 1$	$\{1, \ldots, m\}$
Top Cumulative Probability Intersection (TCPI)	Assume $\boldsymbol{p}, \hat{\boldsymbol{p}}$ sorted in descending order: $m^\star = \min\{m' : \sum_{j=1}^{m'} p_j \geq 0.5\}$ $A_{\boldsymbol{p}} = \{p_j\}_{j=1}^{m^\star}$ $A_{\hat{\boldsymbol{p}}} = \{\hat{p}_j\}_{j=1}^{m^\star}$ $\mathrm{TCPI}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{\lvert A_{\boldsymbol{p}} \cap A_{\hat{\boldsymbol{p}}}\rvert}{m^\star}$	$[0, 1]$

5.4.2 Synthetic Datasets

In this subsection, we evaluate the transition matrix recovery algorithms on multiple synthetic datasets constructed to probe the impact of the following factors on estimation: category imbalance (m), sample size (n), total number of counts (N), and sparsity level of the ground truth stochastic matrix ($\boldsymbol{\Pi} \in \Delta_{m\times d}$). Each factor has three different configurations, and for every combination of the configurations we simulate 10 replicates. This results in a total of 810 datasets.

Experiment setups: In every dataset, we set the number of items from Category 1 as $d = 10$. The number of items from Category 2 (m) takes a value from the set $\{5, 10, 20\}$. For each m, columns of the stochastic matrix ($\boldsymbol{\Pi} \in \Delta_{m\times d}$) are sampled from Dirichlet distribution with parameters choosen from $\{m^{0.5}, m^{-0.5}, m^{-1}\}$[3], see Fig. 5.2 for a heatmap visualization of the simulated stochastic matrices. The number of experiments (n) is selected from $\{5, 10, 40\}$. To introduce variability in the total counts across columns of matrix ($\boldsymbol{X}$), the count (N_i) for the i-th experiment is generated using a negative binomial distribution with a rate parameter 1 and

[3]Dirichlet distribution with parameters ($\boldsymbol{\alpha}$) that are smaller in values has a density function more concentrated around sparse (spiky) distributions.

a probability parameter $\frac{1}{\frac{N_0}{d}+1}$, where N_0 is selected from $\{20, 200, 2000\}$. This results in an expected total column sum of 20, 200, and 2000 respectively.

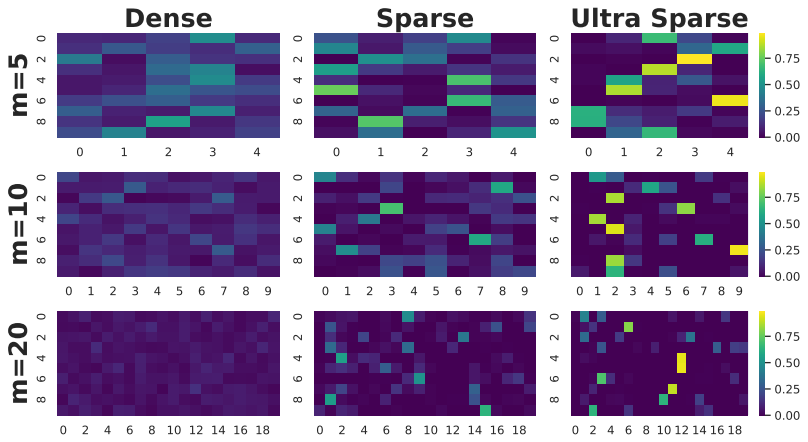
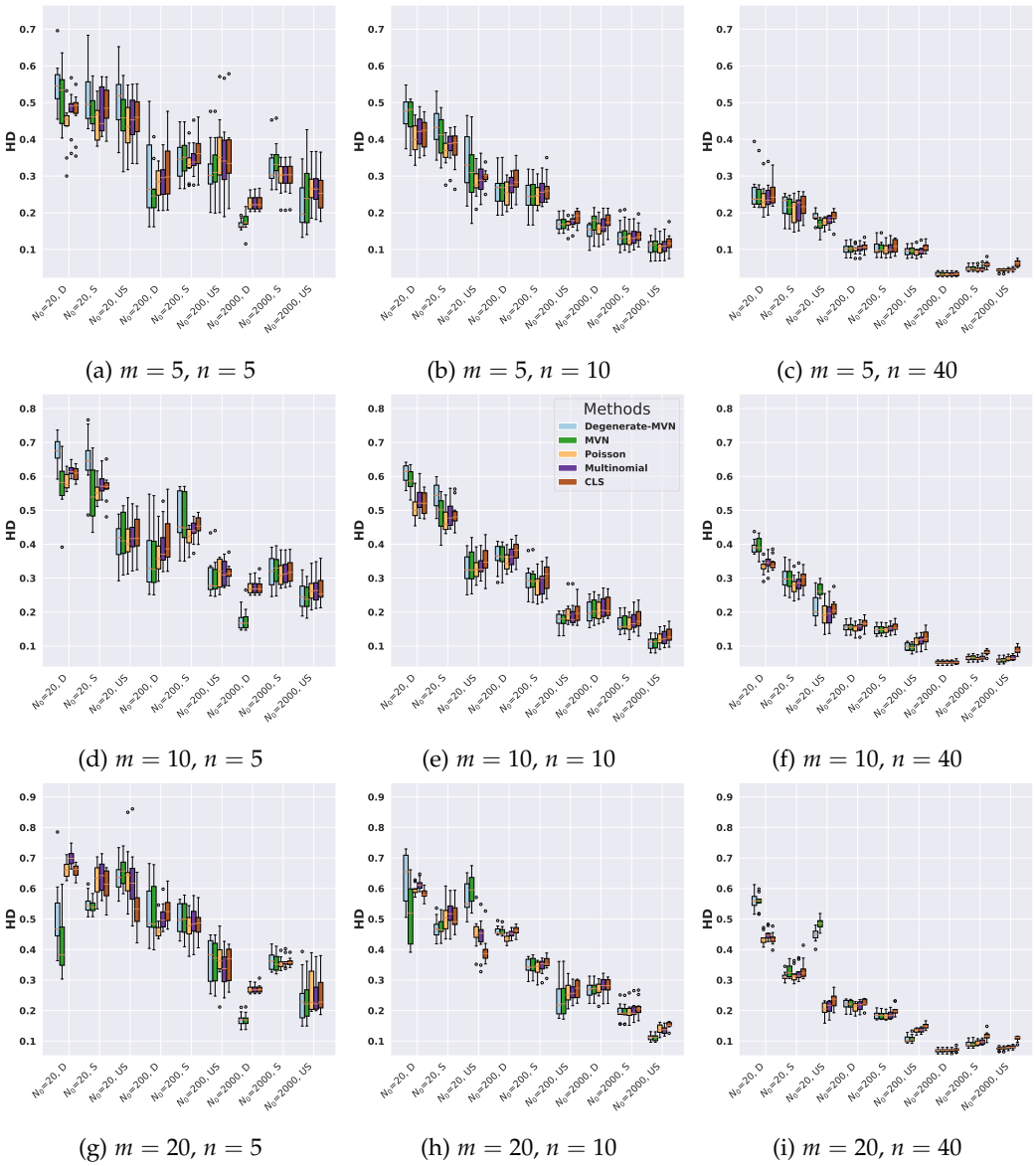


Figure 5.2: Heat map of the simulated stochastic matrix with respect to m (number of items from category 2).

Estimation Results: All the proposed methods based on different approximations from Section 5.2 are evaluated on these synthetic datasets, and the results are summarized in Fig. 5.3 based on Hellinger Distance since its magnitude is invariant to different m, and all metrics share similar trends. Results for rest of the metrics is included in Section 4.6 of the appendix. Our results suggest that number of counts is the most influential factor followed by number of samples. While all the algorithms have similar performances, the two algorithms based on multivariate Gaussian approximations (Section 5.2.3) are the most sensitive to number of counts and sparsity level of the stochastic matrix. The two algorithms signficantly outperforms other methods in experiments with small sample size, high count and dense stochastic matrix, and degrade noticeably in the low count regime.



(a) $m = 5$, $n = 5$ (b) $m = 5$, $n = 10$ (c) $m = 5$, $n = 40$

(d) $m = 10$, $n = 5$ (e) $m = 10$, $n = 10$ (f) $m = 10$, $n = 40$

(g) $m = 20$, $n = 5$ (h) $m = 20$, $n = 10$ (i) $m = 20$, $n = 40$

Figure 5.3: Summary of the prediction performance on all the 81-configurations using Hellinger Distance (HD) as the evaluation metric. HD is bounded between 0 and 1 with 0 means perfect recovery of the ground truth stochastic matrix. Among the four factors considered, number of counts is the most influential factor followed by number of samples. While all the algorithms have similar performance, the two algorithms based on multivariate Gaussian approximations (Section 5.2.3) are the most sensitive to number of counts and sparsity level of the stochastic matrix. The two algorithms outperforms other methods in experiments with small sample size, high count and dense stochastic matrix, and degrade noticeably in the low count regime.

5.4.3 Election Dataset

In New Zealand, electors are apportioned into districts and each voter has two votes where one is for a national party candidate and the other for a local

candidate. This election dataset is publicly available in tabular format collected from each district. We emulate the partial observation exit poll data that only measures marginals for national party selection (Category 1) and local election (Category 2) by summing over the rows and columns of the tabular data. This emulates a real world scenario where the experimenter only has access to the marginals but we have the full ground truth contingency table to evaluate the conditional probabilities recovery accuracies of the proposed methods. The dataset was obtained from R-package [207] and contains election results from 2002, 2005, 2008, 2011, 2014, 2017 and 2020.

Data Preprocessing Since the specific local candidates are distinct for all the districts and vary from year to year, we aggregate the data further by local candidates' party affiliation. Since we expect the voting behavior vary significantly from district to district, while have similar trends across years within the same district, we treat distinct districts as equivalent to different dataset in the synthetic data experiment (i.e they do not share same transition matrix), and different years as different experiments. We retain districts who have data for all years i.e. all 7 elections resulting in a total of 49 districts. In addition, we take the union set of all the parties present in the national elections and local elections across 7 elections to ensure all the transition metrics have same dimension for the purpose of our clustering analysis. The resulting total number of parties present is 55 and 89 respectively for the national elections and the local elections.

Estimation Result We run the proposed methods along with conditional least squares to estimate the stochastic matrix $\mathbf{\Pi}$ for each district across 7 elections.

Note the jk-th entry Π_{jk} will correspond to the probability of a voter voting for party j in the local election given that the voter voted for party k in the national ballot. The evaluations of the algorithms are summarized in Table 5.4.

Clustering Analysis We obtains 49 stochastic matrices which summarize the conditional voting behaviors of each district across the 7 elections. To illustrate that the data has community structure, we perform a clustering analysis on 49 stochastic matrices to identify districts that share similar voting patterns. We visualize the clustering result based on Jensen Shannon divergence in Fig. 5.4, where the rows and columns corresponds to the electorate districts and are ordered based on clustering. The algorithm is able to identity a top cluster includes Te Tai Hauauru, Te Tai Tonga, Waiariki, Tamaki Makaurau and Te Tai Tokerau, and they are the Māori electorates (i.e special electorates that give reserved positions to representatives of New Zealand Parliament). In addition, the two closest districts identified are Bay of Plenty and Tauranga which are both part of the Bay of Plenty Region with similar demographics.

Table 5.4 Summary of the prediction performance on the New Zealand general election dataset. The number in bold means the best performing algorithm for that metric, and the number in parentheses represent standard deviation across 49 distinct districts. Note the first 5 rows of metrics assess the overall agreement, while the last two metrics emphasize on the typical set of the probability mass which are more informative since we are in small sample region with 7 different election years total.

Metric	Degenerate Normal Approximation	Normal Approximation	Multinomial Approximation	Poisson Approximation	CLS
JSD ([0,1])	**0.4233** (0.05422)	0.429 (0.05864)	0.4748 (0.04223)	0.4396 (0.03361)	0.4264 (0.02712)
HD ([0,1])	**0.01344** (0.0009977)	0.01355 (0.001098)	0.0144 (0.0007717)	0.01371 (0.000593)	0.01351 (0.0004604)
MSE ([0,0.05])	0.02264 (0.007572)	0.02436 (0.007907)	0.0307 (0.007618)	0.02051 (0.006633)	**0.01738** (0.006817)
MIRA $\{1,\ldots,89\}$	8.928 (2.008)	**8.652** (2.085)	8.652 (2.153)	10.83 (2.423)	12.44 (2.384)
TCPI [0,1]	0.2474 (0.05938)	**0.2515** (0.06291)	0.234 (0.06144)	0.2006 (0.05974)	0.2007 (0.05318)

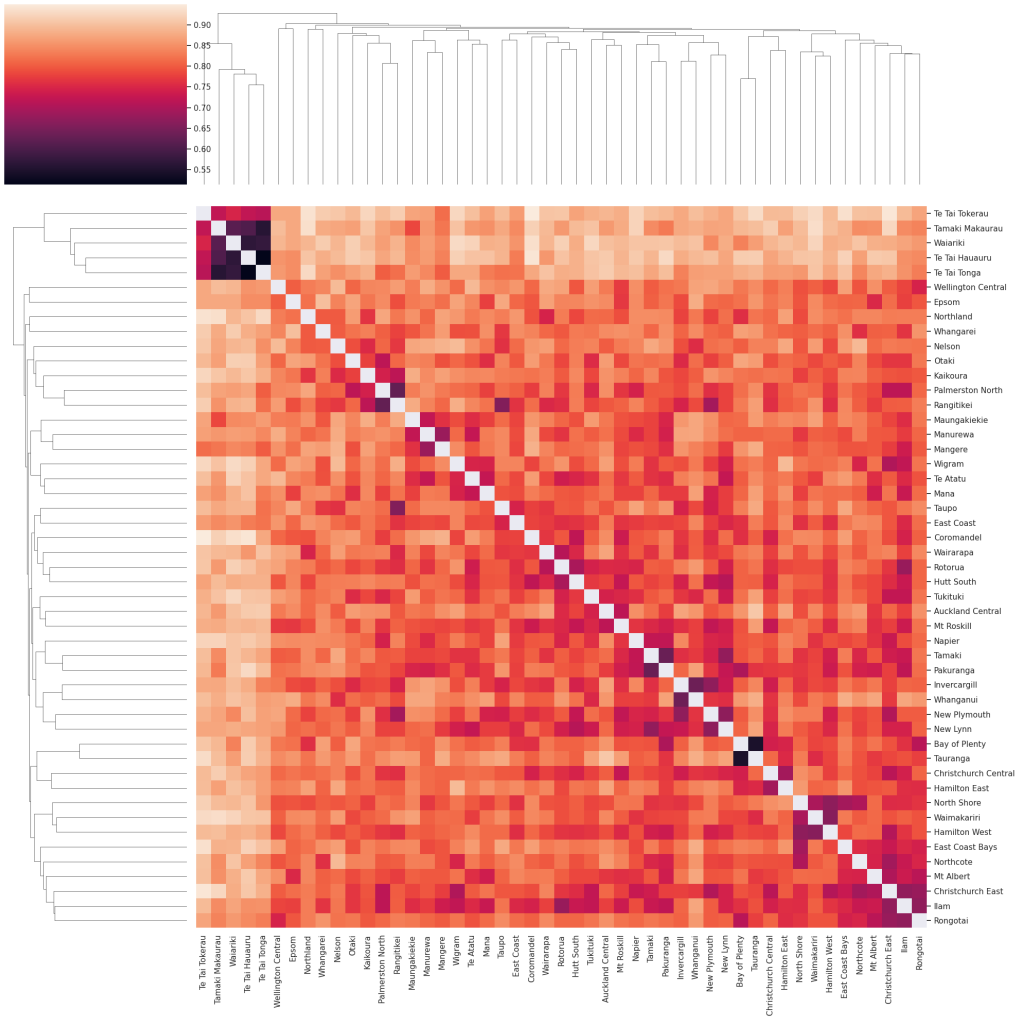


Figure 5.4: Hierarchical Clustering of 49 electoral districts of New Zealand based on the estimated stochastic matrix using Hellinger Distance. Observe the top cluster identity by the algorithm includes Te Tai Hauauru, Te Tai Tonga, Waiariki, Tamaki Makaurau and Te Tai Tokerau are Māori electorates, which are special electorates that give reserved positions to representatives of New Zealand Parliament. In the second cluster, the two closest districts identified are Bay of Plenty and Tauranga which are both part of the Bay of Plenty Region with similar demographics.

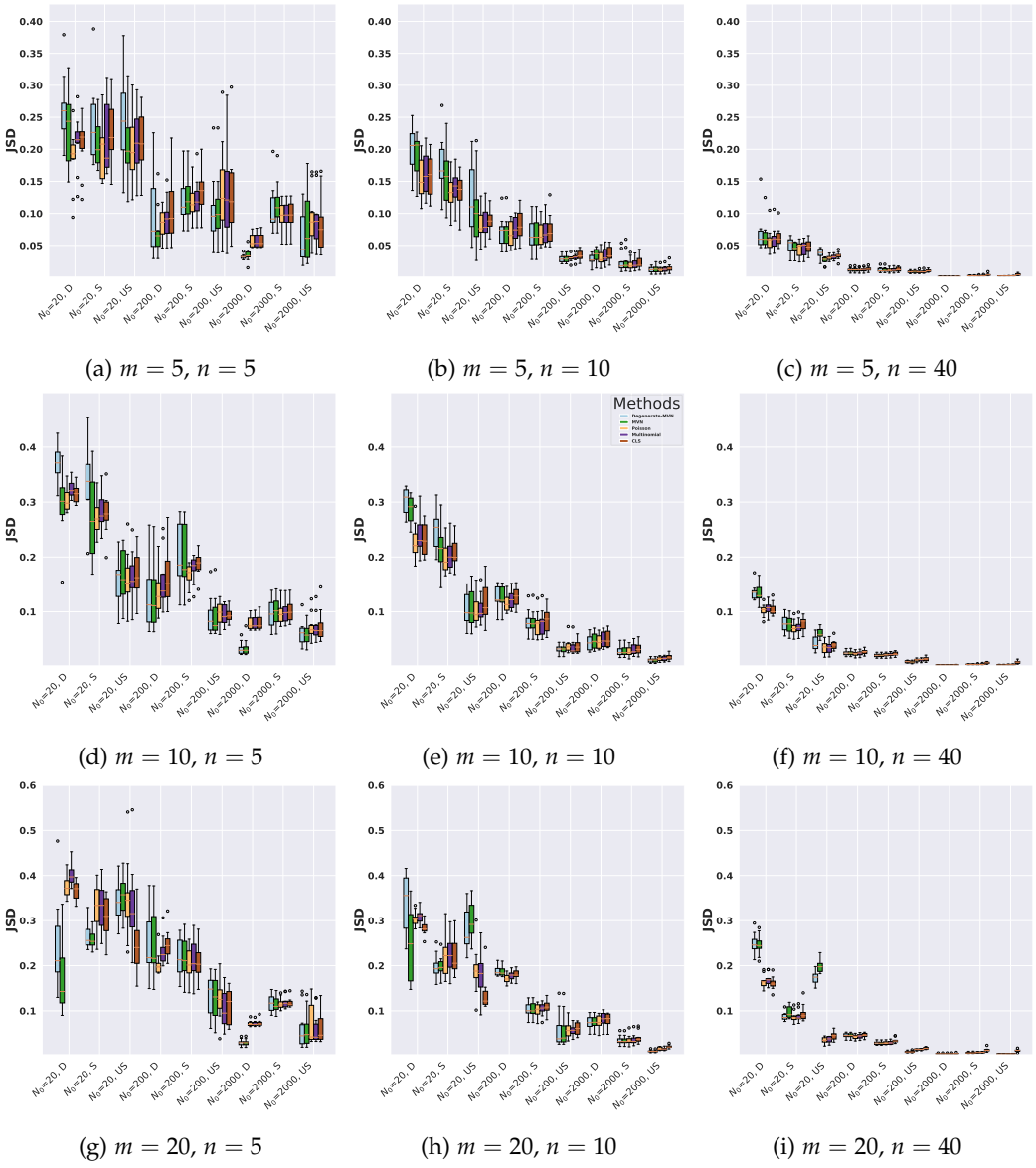
5.5 Conclusion

In this chapter, we propose an exact model with minimal assumptions for the transition recovery problem where only marginals of multiple two-way tabular data are available. To overcome the computational difficulty of the exact model, we provide three valid approximations along with the conditions when they apply.

A novel Riemannian gradient algorithm with Polyak step size is applied to obtain the Maximum Likelihood Estimators (MLE) for the transition matrix. Simulation studies show the scope when those approximation apply. Our experiments on real world dataset from New Zealand general election dataset demonstrate the utility of the method to recover the transition matrix and detect communities within electorate districts based on voting behaviors that are reflective of the demographics.

There are several promising directions for future work. One is to integrate Bayesian frameworks into the proposed model, and this would allow the use of prior knowledge about the structure of the stochastic matrix to enhance the estimation performance and offer uncertainty quantification. Second is to relax the assumption so that the transition matrices of interest have a hierarchical structure that is dependent on experiment index, and this formulation include estimation of state transition matrix from time-inhomogeneous discrete Markov chain as a special case. Another future direction is to study the convergence property of the novel Riemannian gradient algorithm and its extension to a stochastic algorithm. Though the algorithm has shown promising empirical success in this chapter, a rigorous analysis will provide insights on the method and leading to possible extensions and improvements.

5.6 Appendix

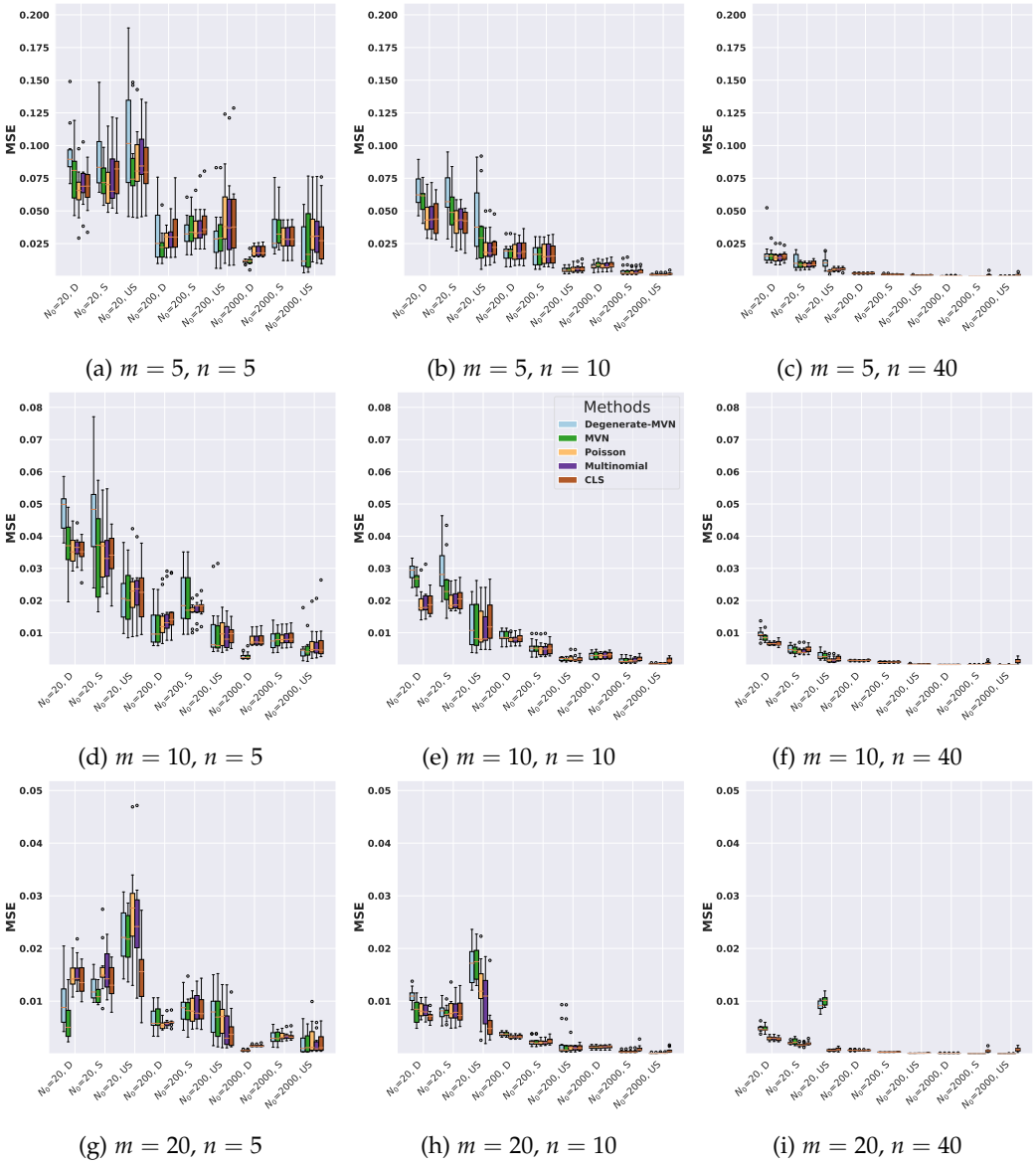


(a) $m = 5$, $n = 5$ (b) $m = 5$, $n = 10$ (c) $m = 5$, $n = 40$

(d) $m = 10$, $n = 5$ (e) $m = 10$, $n = 10$ (f) $m = 10$, $n = 40$

(g) $m = 20$, $n = 5$ (h) $m = 20$, $n = 10$ (i) $m = 20$, $n = 40$

Figure 5.5: Summary of the prediction performance on all the 81-configurations using Jesen-Shannon Divergence (JSD) as the evaluation metric. Jesen-Shannon Divergence is bounded between 0 and 1 with 0 means perfect agreement with the ground truth stochastic matrix.

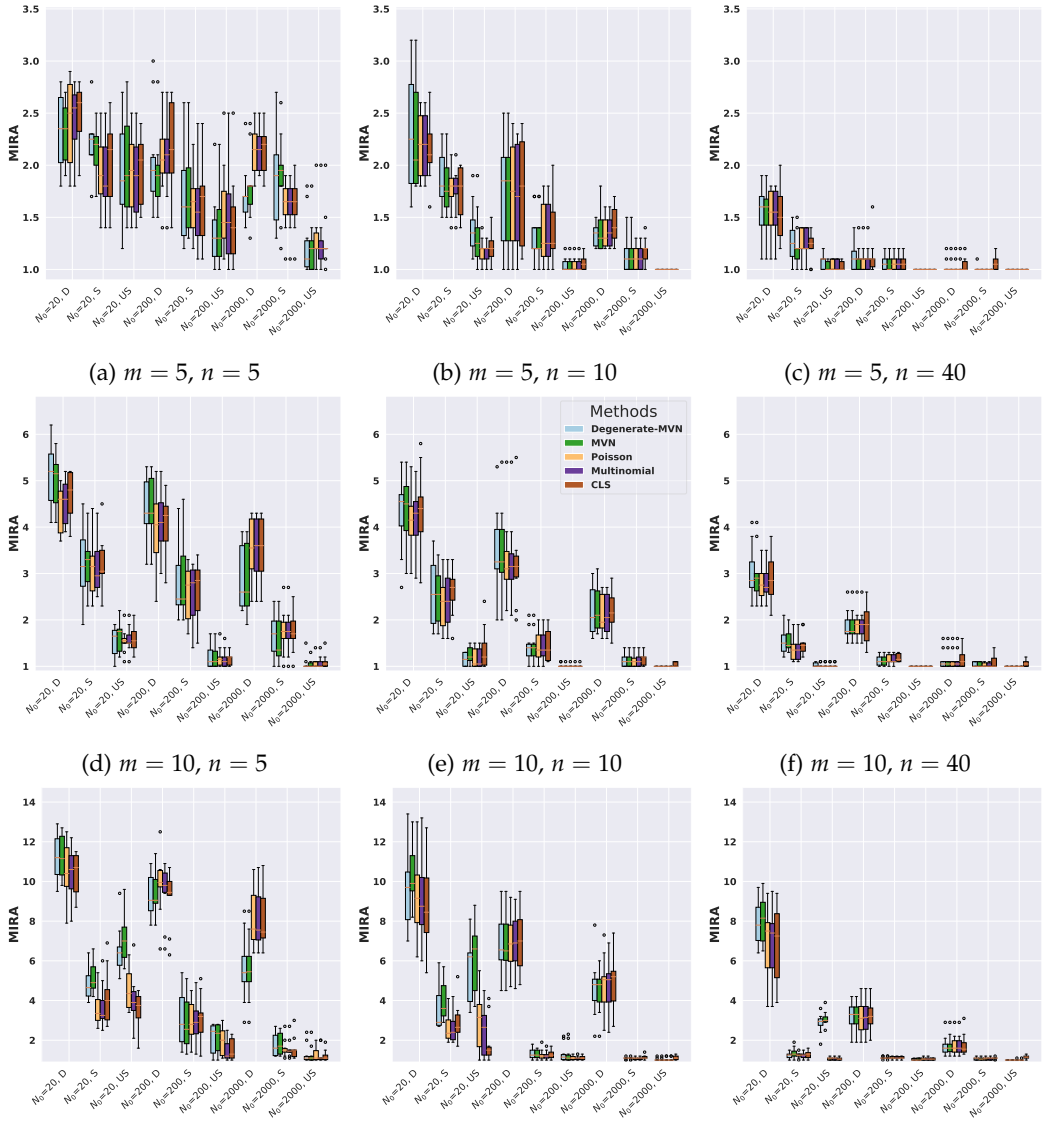


(a) $m = 5$, $n = 5$ (b) $m = 5$, $n = 10$ (c) $m = 5$, $n = 40$

(d) $m = 10$, $n = 5$ (e) $m = 10$, $n = 10$ (f) $m = 10$, $n = 40$

(g) $m = 20$, $n = 5$ (h) $m = 20$, $n = 10$ (i) $m = 20$, $n = 40$

Figure 5.6: Summary of the prediction performance on all the 81-configurations using Mean Square Error (MSE) as the evaluation metric. Mean Square Error is bounded between 0 and $\frac{1}{m}$ with 0 means perfect agreement with the ground truth stochastic matrix.

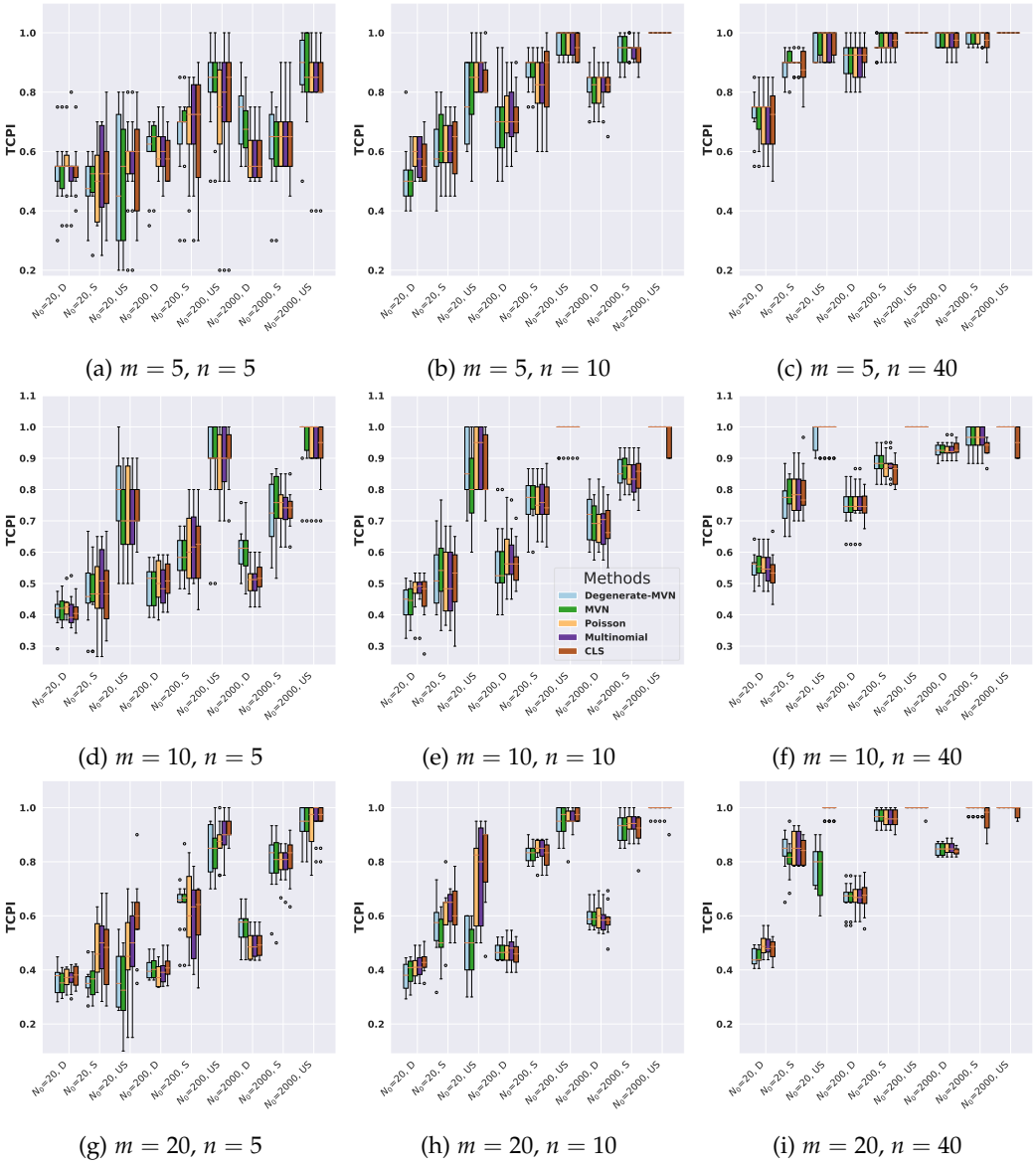


(a) $m = 5$, $n = 5$ (b) $m = 5$, $n = 10$ (c) $m = 5$, $n = 40$

(d) $m = 10$, $n = 5$ (e) $m = 10$, $n = 10$ (f) $m = 10$, $n = 40$

(g) $m = 20$, $n = 5$ (h) $m = 20$, $n = 10$ (i) $m = 20$, $n = 40$

Figure 5.7: Summary of the prediction performance on all the 81-configurations using Maximum Index Rank Agreement (MIRA) as the evaluation metric. Maximum Index Rank Agreement is bounded between 1 and m with 1 means perfectly agreement with the ground truth stochastic matrix in terms of location of the largest entry.



(a) $m = 5$, $n = 5$ (b) $m = 5$, $n = 10$ (c) $m = 5$, $n = 40$

(d) $m = 10$, $n = 5$ (e) $m = 10$, $n = 10$ (f) $m = 10$, $n = 40$

(g) $m = 20$, $n = 5$ (h) $m = 20$, $n = 10$ (i) $m = 20$, $n = 40$

Figure 5.8: Summary of the prediction performance on all the 81-configurations using Top Cumulative Probability Intersection (TCPI) as the evaluation metric. Top Cumulative Probability Intersection is bounded between 0 and $\frac{1}{m}$ with 1 means perfect agreement with the ground truth stochastic matrix in terms of typical set.

CHAPTER VI

Conclusion and Future Work

This thesis focus on methods developed to solve the classical inverse problem and the extend inverse problems arise from real world applications including:

1. a hierarchical Bayesian approach to neutron spectrum unfolding problem
2. a graphical model for fusing diverse microbiome data
3. a hierarchical Bayesian multitask logistic regression model for microbiome Profiling
4. transition matrix recovery problem with application to voting data

All of the chapters in this thesis share a common theme of developing statistical models that taking into account for our understandings of the underlying systems, efficient computational methods and interpretations of our findings.

For the unfolding problem, the present unfolding method could also be coupled to classification algorithms to infer the type and amount of fissile material in unknown neutron sources, for nonproliferation and safeguarding applications. Approximate Bayesian methods can also be investigated for robust unfolding with reduced processing burden.

For fusing diverse microbiome data, one possible future research area is to generalize the model to capture covariance structures of absence-presence datasets

by modeling the binary observations using Bernoulli distributions. Another generalization can be achieved by the incorporation of covariates such as temperature, ph, and physical/chemical perturbations, that may change the composition of the species. The mean of the latent variables can be made a function of the covariates to accomplish that. One another possible area is to incorporate system dynamics into the latent space so as to explicitly capture temporal correlations. In particular, there is increasing interest in collecting longitudinal microbiome data for studying adaptation, resilience, and dynamics over time. The incorporation of a state-space dynamical model into our framework can reveal the temporal evolution of the interactions between the genomes. Another future direction is to improve the parsimony of the model by incorporating sparsity into the latent representation by using sparsity-inducing priors for the covariance or inverse covariance (precision) matrices.

For microbiome Profiling, one future direction is to replace the logit function with other link functions (e.g a probit link function) that have flatter tails so the model is less prone to overconfidence. Second direction is to extend our model to multi-label classification problems, where each task contains multiple binary predictions (e.g diagnosis of different diseases on the same patient). This generalization is of particular interests to the human health prediction application considered in this chapter, since the diseases are not mutually exclusive. Another related extension is to consider the multiclass classification problem, where each task is a classification problem with more than 2 labels (e.g different stages of a disease).

For transition matrix recovery problem, one future direction is to integrate

Bayesian frameworks into the proposed model, and this would allow the use of prior knowledge about the structure of the stochastic matrix to enhance the estimation performance and offer uncertainty quantification. Second is to relax the assumption so that the transition matrices of interest have a hierarchical structure that is dependent on experiment index, and this formulation include estimation of state transition matrix from time-inhomogeneous discrete Markov chain as a special case. Another future direction is to study the convergence property of the novel Riemannian gradient algorithm and its extension to a stochastic algorithm. Though the algorithm has shown promising empirical success in this chapter, a rigorous analysis will provide insights on the method and leading to possible extensions and improvements.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[2] Curtis R Vogel. *Computational methods for inverse problems*. SIAM, 2002.

[3] Haonan Zhu, Yoann Altmann, Angela Di Fulvio, Stephen McLaughlin, Sara Pozzi, and Alfred Hero. A hierarchical bayesian approach to neutron spectrum unfolding with organic scintillators. *IEEE Transactions on Nuclear Science*, 66(10):2265–2274, 2019.

[4] Mehmet Aktukmak, Haonan Zhu, Marc G Chevrette, Julia Nepper, Shruthi Magesh, Jo Handelsman, and Alfred Hero. A graphical model for fusing diverse microbiome data. *arXiv preprint arXiv:2208.09934*, 2022.

[5] John Betteley Birks. *The Theory and Practice of Scintillation Counting*. Pergamon Press, 1964.

[6] K. Weise. An analytical approach to monte carlo spectrum unfolding. *Progress in Nuclear Energy*, 24(1-3):305–310, 1990.

[7] B Wiegel, S Agosteo, R Bedogni, M Caresana, A Esposito, G Fehrenbacher, M Ferrarini, E Hohmann, C Hranitzky, A Kasper, et al. Intercomparison of radiation protection devices in a high-energy stray neutron field, part ii: Bonner sphere spectrometry. *Radiation Measurements*, 44(7-8):660–672, 2009.

[8] Tim Adye. Unfolding algorithms and tests using roounfold. *arXiv preprint arXiv:1105.1160*, 2011.

[9] Chris C Lawrence, Michael Febbraro, Marek Flaska, Sara A Pozzi, and FD Becchetti. Warhead verification as inverse problem: Applications of neutron spectrum unfolding from organic-scintillator measurements. *Journal of Applied Physics*, 120(6):064501, 2016.

[10] SA Pozzi, Yunlin Xu, Thomasz Zak, Shaun D Clarke, Mark Bourne, Marek Flaska, Thomas J Downar, Paolo Peerani, and Vladimir Protopopescu. Fast neutron spectrum unfolding for nuclear non-proliferation and safeguards applications. *Nuovo Cimento. C (Print)*, 33(1):207–214, 2010.

[11] Jeffrey F Williamson, JF Dempsey, AS Kirov, JI Monroe, WR Binns, and Håkan Hedtjärn. Plastic scintillator response to low-energy photons. *Physics in Medicine & Biology*, 44(4):857, 1999.

[12] F.D. Brooks. Development of organic scintillators. *Nuclear Instruments and Methods*, 162(1-3):477–505, 1979.

[13] SA Pozzi, SD Clarke, WJ Walsh, EC Miller, JL Dolan, M Flaska, BM Wieger, A Enqvist, E Padovani, JK Mattingly, et al. Mcnpx-polimi for nuclear nonproliferation applications. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 694:119–125, 2012.

[14] E.C. Miller, S. D. Clarke, M. Flaska, S. Prasad, S.A. Pozzi, and E. Padovani. Mcnpx-polimi post-processing algorithm for detector response simulations. *Journal of Nuclear Materials Management*, XL, 2012.

[15] M.A. Norsworthy, A. Poitrasson-Rivière, M.L. Ruch, S.D. Clarke, and S.A. Pozzi. Evaluation of neutron light output response functions in ej-309 organic scintillators. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 842:20–27, 2017.

[16] Andreas Enqvist, Christopher C. Lawrence, Brian M. Wieger, Sara A. Pozzi, and Thomas N. Massey. Neutron light output response and resolution functions in ej-309 liquid scintillation detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 715:79–86, 2013.

[17] N.V. Kornilov, I. Fabry, S. Oberstedt, and F.-J. Hambsch. Total characterization of neutron detectors with a 252cf source and a new light output determination. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 599(2-3):226–233, 2009.

[18] R Bedogni, G Gualdrini, and F Monteventi. Field parameters and dosimetric characteristics of a fast neutron calibration facility: experimental and monte carlo evaluations. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 476(1):381–385, 2002. Int. Workshop on Neutron Field Spectrometry in Science, Technolog y and Radiation Protection.

[19] Francesco d'Errico, R Ciolini, A Di Fulvio, M Reginatto, J Esposito, C Ceballos Sánchez, and P Colautti. Angle and energy differential neutron spectrometry for the spes bnct facility. *Applied Radiation and Isotopes*, 67(7-8):S141–S144, 2009.

[20] ISO Technical Committee ISO/TC 85/SC 2 Radiological protection. Iso 8529-2:2000. *Reference neutron radiations – Part 2: Calibration fundamentals of radiation protection devices related to the basic quantities characterizing the radiation field*, 2000.

[21] K Weise and M Matzke. A priori distributions from the principle of maximum entropy for the monte carlo unfolding of particle energy spectra. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 280(1):103–112, 1989.

[22] J Pulpan and M Kralik. The unfolding of neutron spectra based on the singular value decomposition of the response matrix. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 325(1-2):314–318, 1993.

[23] Klaus Weise and W Woger. A bayesian theory of measurement uncertainty. *Measurement Science and Technology*, 4(1):1, 1993.

[24] M Matzke and K Weise. Neutron spectrum unfolding by the monte carlo method. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 234(2):324–330, 1985.

[25] Manfred Matzke. Propagation of uncertainties in unfolding procedures. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 476(1–2):230–241, 2002.

[26] Georgios Choudalakis. Fully bayesian unfolding. *arXiv preprint arXiv:1201.4612*, 2012.

[27] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001.

[28] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[29] Mário AT Figueiredo and José M Bioucas-Dias. Restoration of poissonian images using alternating direction optimization. *IEEE Trans. Image Processing*, 19(12):3133–3145, 2010.

[30] Z. T. Harmany, R. F. Marcia, and R. M. Willett. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms - theory and practice. *IEEE Trans. Image Processing*, 21(3):1084–1096, March 2012.

[31] Peter J. Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452, 1990.

[32] B Pehlivanovic, S Avdic, P Marinkovic, SA Pozzi, and M Flaska. Comparison of unfolding approaches for monoenergetic and continuous fast-neutron energy spectra. *Radiation Measurements*, 49:109–114, 2013.

[33] Per Christian Hansen. The l-curve and its use in the numerical treatment of inverse problems. *Computational Inverse Problems in Electrocardiology*, pages 119–142, 2001.

[34] Manfred Matzke. Unfolding of pulse height spectra: the hepro program system. Technical report, SCAN-9501291, 1994.

[35] Marcel Reginatto. The "multi- channel" unfolding programs in the umg package mxd_mc33, grv_mc33, and iqu_mc33 for umg package 3.3 released. 2004.

[36] Manfred Matzke. Unfolding of particle spectra. In *International Conference Neutrons in Research and Industry*, volume 2867, pages 598–608. International Society for Optics and Photonics, 1997.

[37] Marcel Reginatto, Paul Goldhagen, and Sonja Neumann. Spectrum unfolding, sensitivity analysis and propagation of uncertainties with the maximum entropy deconvolution code maxed. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 476(1):242–246, 2002. Int. Workshop on Neutron Field Spectrometry in Science, Technolog y and Radiation Protection.

[38] Mikael Kuusela, Victor M Panaretos, et al. Statistical unfolding of elementary particle spectra: Empirical bayes estimation and bias-corrected uncertainty quantification. *The Annals of Applied Statistics*, 9(3):1671–1705, 2015.

[39] Marcel Reginatto and Andreas Zimbal. Bayesian and maximum entropy methods for fusion diagnostic measurements with compact neutron spectrometers. *Review of Scientific Instruments*, 79(2):023505, 2008.

[40] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

[41] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[42] S. Brooks. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis, 2011.

[43] Y. Altmann, A. Maccarone, A. McCarthy, G. Newstadt, G. S. Buller, S. McLaughlin, and A. Hero. Robust spectral unmixing of sparse multispectral lidar waveforms using gamma Markov random fields. *IEEE Trans. Comput. Imaging*, 3(4):658–670, December 2017.

[44] YongHao Chen, XiMeng Chen, JiaRong Lei, Li An, XiaoDong Zhang, JianXiong Shao, Pu Zheng, and XinHua Wang. Unfolding the fast neutron spectra of a bc501a liquid scintillation detector using gravel method. *Science China Physics, Mechanics & Astronomy*, 57(10):1885–1890, 2014.

[45] Nirmal Keshava and John F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, pages 44–57, January 2002.

[46] M.M. Bourne, S.D. Clarke, M. Paff, A. DiFulvio, M. Norsworthy, and S.A. Pozzi. Digital pile-up rejection for plutonium experiments with solution-grown stilbene. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 842, 2017.

[47] Manfred Matzke. Propagation of uncertainties in unfolding procedures. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 476(1):230–241, 2002. Int. Workshop on Neutron Field Spectrometry in Science, Technolog y and Radiation Protection.

[48] Connor R Fitzpatrick, Isai Salas-González, Jonathan M Conway, Omri M Finkel, Sarah Gilbert, Dor Russ, Paulo José Pereira Lima Teixeira, and Jeffery L Dangl. The plant microbiome: From ecology to reductionism and beyond. *Annual Review of Microbiology*, 74:81–100, 2020.

[49] Gary W Miller. *The exposome: A primer*. Elsevier, 2013.

[50] F James Rohlf and Robert R Sokal. Comparing numerical taxonomic studies. *Systematic Biology*, 30(4):459–490, 1981.

[51] Bob Mau and Michael A Newton. Phylogenetic inference for binary data on dendograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6(1):122–131, 1997.

[52] Vanessa Aguiar-Pulido, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO–S36436, 2016.

[53] Migun Shakya, Chien-Chi Lo, and Patrick SG Chain. Advances and challenges in metatranscriptomic analysis. *Frontiers in genetics*, 10:904, 2019.

[54] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[55] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[56] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[57] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4 of *Information Science and Statistics*. Springer, 2006.

[58] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[59] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[60] Wray Buntine and Aleks Jakulin. Discrete component analysis. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 1–33. Springer, 2005.

[61] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[62] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[63] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[64] David J Harris. Inferring species interactions from co-occurrence data with markov networks. *Ecology*, 97(12):3308–3314, 2016.

[65] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[66] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

[67] Gordana C Popovic, Francis KC Hui, and David I Warton. A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, 165:86–100, 2018.

[68] Gordana C Popovic, David I Warton, Fiona J Thomson, Francis KC Hui, and Angela T Moles. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583, 2019.

[69] Grace Yoon, Irina Gaynanova, and Christian L Müller. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in genetics*, 10:516, 2019.

[70] Bai Zhang and Yue Wang. Learning structural changes of gaussian graphical models in controlled experiments. *arXiv preprint arXiv:1203.3532*, 2012.

[71] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[72] Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

[73] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[74] Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, 17(1):5777–5824, 2016.

[75] Wonyul Lee and Yufeng Liu. Joint estimation of multiple precision matrices with common structures. *The Journal of Machine Learning Research*, 16(1):1035–1062, 2015.

[76] Yuancheng Zhu and Rina Foygel Barber. The log-shift penalty for adaptive estimation of multiple gaussian graphical models. In *Artificial Intelligence and Statistics*, pages 1153–1161. PMLR, 2015.

[77] Bochao Jia and Faming Liang. Joint estimation of multiple mixed graphical models for pan-cancer network analysis. *Stat*, 9(1):e271, 2020.

[78] Xinming Yang, Lingrui Gan, Naveen N Narisetty, and Feng Liang. Gembag: Group estimation of multiple bayesian graphical models. *The Journal of Machine Learning Research*, 22(1):2450–2497, 2021.

[79] Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O'Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.

[80] David I Warton, F Guillaume Blanchet, Robert B O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C Walker, and Francis KC Hui. So many variables: joint modeling in community ecology. *Trends in ecology & evolution*, 30(12):766–779, 2015.

[81] Otso Ovaskainen and Nerea Abrego. *Joint species distribution modelling: with applications in R*. Cambridge University Press, 2020.

[82] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.

[83] Max Welling, Chaitanya Chemudugunta, and Nathan Sutter. Deterministic latent variable models and their pitfalls. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 196–207. SIAM, 2008.

[84] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. *Advances in neural information processing systems*, 21:1089–1096, 2008.

[85] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[86] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[87] Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[88] Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

[89] Mohammad Emtiyaz E Khan, Guillaume Bouchard, Kevin P Murphy, and Benjamin M Marlin. Variational bounds for mixed-data factor analysis. *Advances in Neural Information Processing Systems*, 23:1108–1116, 2010.

[90] Mohammad Khan, Shakir Mohamed, Benjamin Marlin, and Kevin Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Artificial Intelligence and Statistics*, pages 610–618. PMLR, 2012.

[91] Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–473, 2006.

[92] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, A Smith, and M West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742, 2003.

[93] Liang Sun, Shuiwang Ji, Shipeng Yu, and Jieping Ye. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *IJCAI*, volume 9, pages 1230–1235, 2009.

[94] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian exponential family projections for coupled data sources, 2012.

[95] Yasin Yilmaz, Mehmet Aktukmak, and Alfred O Hero. Multimodal data fusion in high-dimensional heterogeneous datasets via generative models. *IEEE Transactions on Signal Processing*, 69:5175–5188, 2021.

[96] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5575–5585, 2018.

[97] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

[98] Yuge Shi, Narayanaswamy Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, pages 15718–15729, 2019.

[99] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

[100] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Representation Learning*, 2019.

[101] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[102] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[103] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[104] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

[105] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv e-prints*, page arXiv:1802.05365, February 2018.

[106] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[107] Akshay Agrawal, Alnur Ali, Stephen Boyd, et al. Minimum-distortion embedding. *Foundations and Trends® in Machine Learning*, 14(3):211–378, 2021.

[108] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[109] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[110] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[111] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[112] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[113] Iain Murray Zoubin Ghahramani. A note on the evidence and bayesian occam's razor. Technical report, Gatsby Unit Technical Report, 2005.

[114] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[115] Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67, 2004.

[116] Cédric Archambeau and Francis Bach. Sparse probabilistic projections. *Advances in neural information processing systems*, 21, 2008.

[117] Magnus Rattray, Oliver Stegle, Kevin Sharp, and John Winn. Inference algorithms and learning theory for bayesian sparse factor analysis. In *Journal of Physics: Conference Series*, volume 197, page 012002. IOP Publishing, 2009.

[118] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[119] Sadanori Konishi and Genshiro Kitagawa. Information criteria and statistical modeling. *Springer*, 2008.

[120] Matthew J Beal and Zoubin Ghahramani. Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–831, 2006.

[121] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[122] Amanda Hurley, Marc G. Chevrette, Natalia Rosario-Meléndez, Jo Handelsman, and Gerard D. Wright. Thor's hammer: the antibiotic koreenceine drives gene expression in a model microbial community. *mBio*, 0(0):e02486–21, 2022.

[123] Johan Bengtsson-Palme. Microbial model communities: To understand complexity, harness the power of simplicity. *Computational and Structural Biotechnology Journal*, 18:3987–4001, 2020.

[124] Karen De Roy, Massimo Marzorati, Pieter Van den Abbeele, Tom Van de Wiele, and Nico Boon. Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental Microbiology*, 16(6):1472–1481, 2014.

[125] Marc G. Chevrette, Jennifer R. Bratburd, Cameron R. Currie, Reed M. Stubbendieck, and Barbara Methe. Experimental microbiomes: Models not to scale. *mSystems*, 4(4):e00175–19, 2019.

[126] Benjamin E. Wolfe. Using cultivated microbial communities to dissect microbiome assembly: Challenges, limitations, and the path ahead. *mSystems*, 3(2):e00161–17, 2018.

[127] Gabriel L. Lozano, Juan I. Bravo, Manuel F. Garavito Diago, Hyun Bong Park, Amanda Hurley, S. Brook Peterson, Eric V. Stabb, Jason M. Crawford, Nichole A. Broderick, Jo Handelsman, Gary M. Dunny, Roberto Kolter, and Irene Newton. Introducing thor, a model microbiome for genetic dissection of community behavior. *mBio*, 10(2):e02846–18, 2019.

[128] Gabriel L. Lozano, Hyun Bong Park, Juan I. Bravo, Eric A. Armstrong, John M. Denu, Eric V. Stabb, Nichole A. Broderick, Jason M. Crawford, Jo Handelsman, and Marie A. Elliot. Bacterial analogs of plant tetrahydropyridine alkaloids mediate microbial interactions in a rhizosphere model system. *Applied and Environmental Microbiology*, 85(10):e03058–18, 2019.

[129] Franziska Liesecke, Dimitri Daudu, Rodolphe Dugé de Bernonville, Sébastien Besseau, Marc Clastre, Vincent Courdavault, Johan-Owen De Craene, Joel Crèche, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, et al. Ranking genome-wide correlation measurements improves microarray and rna-seq based global and targeted co-expression networks. *Scientific reports*, 8(1):1–16, 2018.

[130] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[131] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[132] Xiaogang Wang, Cha Zhang, and Zhengyou Zhang. Boosted multi-task learning for face verification with applications to web image and video search. In *2009 IEEE conference on computer vision and pattern recognition*, pages 142–149. IEEE, 2009.

[133] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[134] Christian Widmer and Gunnar Rätsch. Multitask learning in computational biology. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 207–216. JMLR Workshop and Conference Proceedings, 2012.

[135] Priyadip Ray, Lingling Zheng, Joseph Lucas, and Lawrence Carin. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376, 2014.

[136] Lora V Hooper, Tore Midtvedt, and Jeffrey I Gordon. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual review of nutrition*, 22:283, 2002.

[137] Fredrik Backhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920, 2005.

[138] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[139] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[140] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017.

[141] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[142] Yi-Hui Zhou and Paul Gallins. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in genetics*, 10:579, 2019.

[143] Sachin Aryal, Ahmad Alimadadi, Ishan Manandhar, Bina Joe, and Xi Cheng. Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertension*, 76(5):1555–1562, 2020.

[144] Giovanni Cammarota, Gianluca Ianiro, Anna Ahern, Carmine Carbone, Andriy Temko, Marcus J Claesson, Antonio Gasbarrini, and Giampaolo Tortora. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews gastroenterology & hepatology*, 17(10):635–648, 2020.

[145] Ishan Manandhar, Ahmad Alimadadi, Sachin Aryal, Patricia B Munroe, Bina Joe, and Xi Cheng. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 320(3):G328–G337, 2021.

[146] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[147] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[148] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. CRC press, 2018.

[149] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348, 2009.

[150] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. *Advances in neural information processing systems*, 23, 2010.

[151] André R Gonçalves, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning with gaussian copula models. *The Journal of Machine Learning Research*, 17(1):1205–1234, 2016.

[152] Shengbo Guo, Onno Zoeter, and Cédric Archambeau. Sparse bayesian multi-task learning. *Advances in Neural Information Processing Systems*, 24, 2011.

[153] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(1), 2007.

[154] Yi Zhang and Jeff Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in neural information processing systems*, 23, 2010.

[155] Andre Goncalves, Priyadip Ray, Braden Soper, David Widemann, Mari Nygård, Jan F Nygård, and Ana Paula Sales. Bayesian multitask learning regression for heterogeneous patient cohorts. *Journal of Biomedical Informatics*, 100:100059, 2019.

[156] Cécile Bazot, Nicolas Dobigeon, Jean-Yves Tourneret, and Alfred O Hero. A bernoulli-gaussian model for gene factor analysis. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5996–5999. IEEE, 2011.

[157] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[158] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

[159] Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, 2014.

[160] Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications*, 6(1):6528, 2015.

[161] Krithivasan Sankaranarayanan, Andrew T Ozga, Christina Warinner, Raul Y Tito, Alexandra J Obregon-Tito, Jiawu Xu, Patrick M Gaffney, Lori L Jervis, Derrell Cox, Lancer Stephens, et al. Gut microbiome diversity among cheyenne and arapaho individuals from western oklahoma. *Current Biology*, 25(24):3161–3169, 2015.

[162] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, 2017.

[163] Anna Heintz-Buschart, Patrick May, Cédric C Laczny, Laura A Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, Jochen G Schneider, Angela Hogan, Carine De Beaufort, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology*, 2(1):1–13, 2016.

[164] Emily Vogtmann, Xing Hua, Georg Zeller, Shinichi Sunagawa, Anita Y Voigt, Rajna Hercog, James J Goedert, Jianxin Shi, Peer Bork, and Rashmi Sinha. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PloS one*, 11(5):e0155362, 2016.

[165] Zhuye Jie, Huihua Xia, Shi-Long Zhong, Qiang Feng, Shenghui Li, Suisha Liang, Huanzi Zhong, Zhipeng Liu, Yuan Gao, Hui Zhao, et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nature communications*, 8(1):845, 2017.

[166] Jing Li, Fangqing Zhao, Yidan Wang, Junru Chen, Jie Tao, Gang Tian, Shouling Wu, Wenbin Liu, Qinghua Cui, Bin Geng, et al. Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, 5:1–19, 2017.

[167] Rohit Loomba, Victor Seguritan, Weizhong Li, Tao Long, Niels Klitgord, Archana Bhatt, Parambir Singh Dulai, Cyrielle Caussy, Richele Bettencourt, Sarah K Highlander, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell metabolism*, 30(3):607, 2019.

[168] Dorottya Nagy-Szakal, Brent L Williams, Nischay Mishra, Xiaoyu Che, Bohyun Lee, Lucinda Bateman, Nancy G Klimas, Anthony L Komaroff, Susan Levine, Jose G Montoya, et al. Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome*, 5(1):1–17, 2017.

[169] Gavin M Douglas, Richard Hansen, Casey MA Jones, Katherine A Dunn, André M Comeau, Joseph P Bielawski, Rachel Tayler, Emad M El-Omar, Richard K Russell, Georgina L Hold, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric crohn's disease. *Microbiome*, 6:1–12, 2018.

[170] Silas Kieser, Shafiqul A Sarker, Olga Sakwinska, Francis Foata, Shamima Sultana, Zeenat Khan, Shoheb Islam, Nadine Porta, Séverine Combremont, Bertrand Betrisey, et al. Bangladeshi children with acute diarrhoea show faecal microbiomes with increased streptococcus abundance, irrespective of diarrhoea aetiology. *Environmental microbiology*, 20(6):2256–2269, 2018.

[171] Bruce A Rosa, Taniawati Supali, Lincoln Gankpala, Yenny Djuardi, Erliyani Sartono, Yanjiao Zhou, Kerstin Fischer, John Martin, Rahul Tyagi, Fatorma K Bolay, et al. Differential human gut microbiome assemblages during soil-transmitted helminth infections in indonesia and liberia. *Microbiome*, 6(1):1–19, 2018.

[172] Zi Ye, Ni Zhang, Chunyan Wu, Xinyuan Zhang, Qingfeng Wang, Xinyue Huang, Liping Du, Qingfeng Cao, Jihong Tang, Chunjiang Zhou, et al. A metagenomic study of the gut microbiome in behcet's disease. *Microbiome*, 6(1):1–13, 2018.

[173] Paolo Ghensi, Paolo Manghi, Moreno Zolfo, Federica Armanini, Edoardo Pasolli, Mattia Bolzan, Alberto Bertelle, Federico Dell'Acqua, Ester Dellasega, Romina Waldner, et al. Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. *npj Biofilms and Microbiomes*, 6(1):47, 2020.

[174] Meagan A Rubel, Arwa Abbas, Louis J Taylor, Andrew Connell, Ceylan Tanes, Kyle Bittinger, Valantine N Ndze, Julius Y Fonsah, Eric Ngwang, André Essiane, et al. Lifestyle and the presence of helminths is associated with gut microbiome composition in cameroonians. *Genome biology*, 21(1):1–32, 2020.

[175] Feng Zhu, Yanmei Ju, Wei Wang, Qi Wang, Ruijin Guo, Qingyan Ma, Qiang Sun, Yajuan Fan, Yuying Xie, Zai Yang, et al. Metagenome-wide association of gut microbiome features for schizophrenia. *Nature communications*, 11(1):1612, 2020.

[176] LM Proctor, HH Creasy, JM Fettweis, J Lloyd-Price, A Mahurkar, W Zhou, GA Buck, MP Snyder, JF Strauss, GM Weinstock, et al. The integrative hmp (ihmp) research network consortium. *The integrative human microbiome project Nature*, 569(7758):641–648, 2019.

[177] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[178] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[179] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[180] Baohong Wang, Mingfei Yao, Longxian Lv, Zongxin Ling, and Lanjuan Li. The human microbiota in health and disease. *Engineering*, 3(1):71–82, 2017.

[181] Charles Soussen, Jérôme Idier, David Brie, and Junbo Duan. From bernoulli–gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, 2011.

[182] Dankmar Böhning and Bruce G Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

[183] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[184] Tommi S Jaakkola and Michael I Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, pages 283–294. PMLR, 1997.

[185] Jan Drugowitsch. Variational bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*, 2013.

[186] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[187] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[188] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[189] Liping Liu, Daniel Sheldon, and Thomas Dietterich. Gaussian approximation of collective graphical models. In *International Conference on Machine Learning*, pages 1602–1610. PMLR, 2014.

[190] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *ACM computing surveys (CSUR)*, 51(4):1–36, 2018.

[191] Gary King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press, 1997.

[192] Seth R Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported obama in 2012? ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298, 2015.

[193] Matthew T Jones. Estimating markov transition matrices using proportions data: an application to credit risk. 2005.

[194] Garrett Bernstein and Daniel Sheldon. Consistently estimating markov chains with noisy aggregate data. In *Artificial intelligence and statistics*, pages 1142–1150. PMLR, 2016.

[195] George A Miller. Finite markov processes in psychology. *Psychometrika*, 17(2):149–167, 1952.

[196] Albert Madansky. Least squares estimation in finite markov processes. *Psychometrika*, 24(2):137–144, 1959.

[197] Dennis J Aigner and Stephen M Goldfeld. Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica: Journal of the Econometric Society*, pages 113–134, 1974.

[198] Adriaan P Van Der Plas. On the estimation of the parameters of markov probability models using macro data. *The Annals of Statistics*, 11(1):78–85, 1983.

[199] John David Kalbfleisch, Jerald Franklin Lawless, and William M Vollmer. Estimation in markov models from aggregate data. *Biometrics*, pages 907–919, 1983.

[200] Rahul Singh, Isabel Haasler, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Inference with aggregate data in probabilistic graphical models: An optimal transport approach. *IEEE Transactions on Automatic Control*, 67(9):4483–4497, 2022.

[201] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010.

[202] André Klima, Paul W Thurner, Christoph Molnar, Thomas Schlesinger, and Helmut Küchenhoff. Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches. *AStA Advances in Statistical Analysis*, 100:133–159, 2016.

[203] André Klima, Thomas Schlesinger, Paul W Thurner, and Helmut Küchenhoff. Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research*, 48(2):296–325, 2019.

[204] Rafael Romero, Jose M Pavía, Jorge Martín, and Gerardo Romero. Assessing uncertainty of voter transitions estimated from aggregated data. application to the 2017 french presidential election. *Journal of Applied Statistics*, 47(13-15):2711–2736, 2020.

[205] Karl Pearson. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika*, 11(3):145–158, 1916.

[206] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[207] Jose M Pavía. ei. datasets: real data sets for assessing ecological inference algorithms. *Social Science Computer Review*, 40(1):247–260, 2022.

[208] Bero Roos. Metric multivariate poisson approximation of the generalized multinomial distribution. *Theory of Probability & Its Applications*, 43(2):306–316, 1999.

[209] Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the structure, covering, and learning of poisson multinomial distributions. In *2015 IEEE 56th annual symposium on foundations of computer science*, pages 1203–1217. IEEE, 2015.

[210] Zhengzhi Lin, Yueyao Wang, and Yili Hong. The poisson multinomial distribution and its applications in voting theory, ecological inference, and machine learning. *arXiv preprint arXiv:2201.04237*, 2022.

[211] Yves Grandvalet and Yoshua Bengio. Entropy regularization., 2006.

[212] John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

[213] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[214] Andrew D Barbour. Stein's method and poisson process convergence. *Journal of Applied Probability*, 25(A):175–184, 1988.

[215] Paul Deheuvels and Dietmar Pfeifer. Poisson approximations of multinomial distributions and point processes. *Journal of multivariate analysis*, 25(1):65–89, 1988.

[216] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1:32, 1987.

[217] Elad Hazan and Sham Kakade. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

[218] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[219] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[220] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 2020.

[221] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

[222] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *lecture notes of EE3920, Stanford University, Autumn Quarter*, 2004:2004–2005, 2003.

[223] Ryan D'Orazio, Nicolas Loizou, Issam Laradji, and Ioannis Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *arXiv preprint arXiv:2110.15412*, 2021.

[224] Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

[225] Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.

[226] Andrew Holbrook. Differentiating the pseudo determinant. *Linear Algebra and its Applications*, 548:293–304, 2018.