The Phylogenetic LASSO and the Microbiome: Metagenomic Modeling in Fecal Microbiota Transplantation

by

Stephen T. A. Rush

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Mathematics

Guelph, Ontario, Canada
©Stephen T. A. Rush, December, 2017
ABSTRACT

THE PHYLOGENETIC LASSO AND THE MICROBIOME: METAGENOMIC MODELING IN FECAL MICROBIOTA TRANSPLANTATION

Stephen T. A. Rush
University of Guelph, 2017

Advisors:
Peter T. Kim
Rajesh Pereira

Advisory Committee:
J. Scott Weese

Scientific investigations that incorporate next generation sequencing involve analyses of high-dimensional data where the need to organize, collate and interpret the outcomes are pressingly important. Data of the microbiome can currently be collected, leading to possible advances in personalized medicine. In this thesis, we lay down a statistical framework for incorporating metagenomic information in predictive modeling with a view toward synthesis of products tailored to individual patients. In particular, we develop the phylogenetic LASSO ($\Phi$-LASSO), a form of model regularization which incorporates known
relationships between predictors for the purpose of model selection. We apply the \( \Phi \)-LASSO to a pilot metagenomic study on the efficacy of fecal microbiota transplantation in the treatment of *Clostridium difficile* infections. Although the thesis applies the technique to data for a particular infectious disease, the methodology is sufficiently rich to be expanded to other problems in medicine, especially those in which coincident ‘-omics’ covariates and clinical responses are simultaneously captured.
ACKNOWLEDGEMENTS

I thank my advisors, Drs Peter Kim and Rajesh Pereira; committee member Dr Scott Weese, whose lab space we borrowed for sequencing; and Dr Christine Lee of St. Joseph’s Healthcare Hamilton, on whose clinical work we base our application.

The Clostridium difficile infection study was approved by

(i) St. Joseph’s Healthcare and McMaster University REB 05-2477

(ii) University of Guelph REB 12AU013

All laboratory procedures were performed in the Department of Pathology and Molecular Medicine, St Joseph’s Healthcare Hamilton, and the Department of Pathobiology, University of Guelph. Funding was provided by grants from CIHR, NSERC and St Joseph’s Healthcare Hamilton.

Finally, I would like to thank friends and family who supported me along the way.
# Table of Contents

Abstract

Acknowledgements

List of Tables

List of Figures

Abbreviations

Symbols

1 Introduction

2 Model Selection
   2.1 Models and estimators .................................................. 6
   2.2 Generalized linear models ............................................. 7
   2.3 Competing interests .................................................... 9
      2.3.1 Estimation and Risk .............................................. 11
      2.3.2 Model Selection Consistency ................................... 12
   2.4 Regularized Maximum Likelihood ..................................... 14
      2.4.1 Convexity .......................................................... 16
      2.4.2 Structured sparsity and group selection ....................... 19

3 Metagenomics
   3.1 Information System .................................................... 22
   3.2 A Bacterial Discriminant .............................................. 26
      3.2.1 Species proxies .................................................. 28
   3.3 Defining Bacteria ...................................................... 29
      3.3.1 Diversity in Propagation ....................................... 29
      3.3.2 Distributional bacteria ......................................... 32
      3.3.3 Surveying Bacterial Communities ............................... 34
4 The phylogenetic LASSO 36
4.1 Generalized linear models and the \( \Phi \)-LASSO 37
4.1.1 The \( \Phi \)-LASSO 38
4.1.2 \( \Phi \)-LASSO algorithm 42
4.2 Simulations 43
4.2.1 Simulation 1 43
4.2.2 Simulation 2 47
4.2.3 Additional simulations 49
4.3 Theoretical Results 52
4.3.1 LAN conditions and the Oracle 54
4.4 Proofs 57
4.4.1 Proof of Theorem 4.3.5 60
4.4.2 Proof of Theorem 4.3.6 63
4.4.3 Proof of Theorem 4.3.7 67

5 Clostridium difficile: A public health concern 69
5.1 Background 69
5.2 Application 71
5.2.1 Description 71
5.2.2 Analysis 73

6 Conclusion 77

Bibliography 80

A \( \Phi \)-LASSO source code 95
A.0.1 R-code 95
A.0.2 fortran-code 99
List of Tables

2.1 Median estimation error (SSE) for Gaussian model $Y = X\beta + \epsilon$ where $X \in \mathbb{R}^{50 \times 40}$ with entries drawn from the standard normal, with $\epsilon \sim N(0, \sigma_e)$, based on 100 replicates. The interquartile range is provided in parentheses. Simulation described in Section 2.3. .................................................. 11

2.2 Median prediction error (SSPE) for Gaussian model $Y = X\beta + \epsilon$ where $X \in \mathbb{R}^{50 \times 40}$ with entries drawn from the standard normal, with $\epsilon \sim N(0, \sigma_e)$, based on 100 replicates. The interquartile range is provided in parentheses. Simulation described in Section 2.3. .................................................. 11

4.1 An example taxonomy generated for thirteen OTUs using five taxon levels. ........................................... 40

4.2 Simulation 1. Estimation and prediction error for the oracle estimator (OLS), $\Phi$-LASSO, and SCAD. Presented are the mean error (standard error). .................................................. 51

4.3 Simulation 1. Recall and precision for the $\Phi$-LASSO and SCAD. Presented are the mean (standard error). .................................................. 51

4.4 Simulation 2. Estimation and prediction error for the oracle estimator (OLS), $\Phi$-LASSO, and SCAD. Presented are the mean error (standard error). .................................................. 51

4.5 Simulation 2. Recall and precision for the $\Phi$-LASSO and SCAD. Presented are the mean (standard error). .................................................. 52

5.1 The selected OTUs for pre-FMT microbiomes at the tuning parameters selected by AUC ($\lambda = 0.0049$) and BS ($\lambda = 0.0067$) leave-one-out cross-validation on relative abundances. .................................................. 75

5.2 The selected OTUs for post-FMT microbiomes at the tuning parameters selected by AUC ($\lambda = 0.0065$) and BS ($\lambda = 0.0056$) leave-one-out cross-validation on relative abundances. .................................................. 76
List of Figures

1.1 Phylogenetic tree. The branching structure indicates proximity of evolutionary relationship. .................................................. 3

3.1 (a) Distribution of DNA sequences over sequence space $\mathcal{S}$. The dotted circles surround clusters representing four separate lineages. (b) Underlying phylogeny of distribution presented in (a). ............................................. 25

3.2 Four clusters of points from classes ‘o’, ‘x’, ‘c’, and ‘s’ sampled uniformly from unit disks centred at (1,2), (3,3), (3.5,1), and (5,1), respectively, with $n = 20$ points in each class. See Section 3.2.1. .................................................. 30

3.3 Depiction of three different hierarchical clustering methods applied to the points in Figure 3.2. Employed are the complete, average, and single linkage methods with Euclidean, maximum, and Manhattan metrics. See Section 3.2.1. ............................................. 31

4.1 A graphical representation of the taxonomy as well as an illustration of the decomposition of the parameters $\beta$ following the taxonomy in Table 4.1. .................................................. 41

4.2 Heatmap of covariance matrix for 10,000 point validation set from the tuning parameter step. Displayed is the submatrix corresponding to a single ‘order’. .................................................. 44

4.3 Pruned taxonomy for simulations 1 and 2, from an initial 4096 and 16384 leaves, respectively. .................................................. 45

4.4 Simulation 1. Median recall (solid) and precision (dashed) for $\Phi$-LASSO against log($\lambda$) for sample sizes $n = 50, 100, 150$. Included is the median MSPE curve scaled by largest median (dotted). The ‘+’ indicates the chosen (log) tuning parameter. .................................................. 50

5.1 AUC and BS obtained by leave-one-out cross-validation for pre-FMT OTUs. Labeled are the tuning parameters by (a) AUC ($\lambda = 0.0049$) and (b) BS ($\lambda = 0.0067$). .................................................. 75

5.2 AUC and BS obtained by leave-one-out cross-validation for post-FMT OTUs. Labeled are the tuning parameters by (a) AUC ($\lambda = 0.0065$) and (b) BS ($\lambda = 0.0056$). .................................................. 76
ABBREVIATIONS

AUC  area under the curve
bp   base pair
BS   Brier score
CAP  composite absolute penalty
CCCP concave-convex procedure
CDI  Clostridium difficile infection
CF   cell fusion
FMT  fecal microbiota transplantation
GLM  generalized linear model
HGT  horizontal gene transfer
H-LASSO hierarchical LASSO
KKT  Karush-Kuhn-Tucker
LAN  local asymptotic normality
LASSO least absolute shrinkage and selection operator
LLA  local linear approximation
LOO-CV leave-one-out cross-validation
MCP  minimax concave penalty
MPS  massively parallel sequencing
MSPE mean squared prediction error
NGS  next-generation sequencing
OLS  ordinary least squares
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU</td>
<td>operational taxonomical unit</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>Φ-LASSO</td>
<td>phylogenetic LASSO</td>
</tr>
<tr>
<td>(r)-MLE</td>
<td>(regularized) maximum likelihood estimate</td>
</tr>
<tr>
<td>SCAD</td>
<td>smoothly clipped absolute deviation</td>
</tr>
<tr>
<td>SSPE</td>
<td>sum of squared prediction error</td>
</tr>
<tr>
<td>SSE</td>
<td>sum of squared estimation error</td>
</tr>
</tbody>
</table>
## SYMBOLS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{A}_n^0, \hat{\mathcal{A}}_n )</td>
<td>index set of true parameters, index set of non-zero estimates</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>individual effects coefficients</td>
</tr>
<tr>
<td>( \beta, \beta^0, \beta_n^0 )</td>
<td>parameter vector</td>
</tr>
<tr>
<td>( \hat{\beta}, \hat{\beta}_n )</td>
<td>parameter estimate</td>
</tr>
<tr>
<td>( \chi )</td>
<td>sample space</td>
</tr>
<tr>
<td>( d, D )</td>
<td>groups coefficients</td>
</tr>
<tr>
<td>( E )</td>
<td>evolutionary space</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>error</td>
</tr>
<tr>
<td>( f )</td>
<td>probability density function</td>
</tr>
<tr>
<td>( F )</td>
<td>probability distribution function</td>
</tr>
<tr>
<td>( g )</td>
<td>link function</td>
</tr>
<tr>
<td>( \mathcal{I}, \mathcal{I}(\beta) )</td>
<td>Fisher information matrix</td>
</tr>
<tr>
<td>( \ell )</td>
<td>log-likelihood function</td>
</tr>
<tr>
<td>( L )</td>
<td>loss function (Chapter 2), lineage (Chapter 4)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>tuning parameter (vector)</td>
</tr>
<tr>
<td>( \eta )</td>
<td>linear predictor</td>
</tr>
<tr>
<td>( n )</td>
<td>sample size</td>
</tr>
<tr>
<td>( N )</td>
<td>normal distribution</td>
</tr>
<tr>
<td>( N )</td>
<td>set of positive integers</td>
</tr>
<tr>
<td>( o, o_p )</td>
<td>little ‘oh’, little ‘oh’ in probability</td>
</tr>
<tr>
<td>( O, O_p )</td>
<td>big ‘oh’, big ‘oh’ in probability</td>
</tr>
</tbody>
</table>
\( p, p_n \) \hspace{1cm} \text{number of parameters}

\( p_\lambda \) \hspace{1cm} \text{penalty function}

\( \varphi \) \hspace{1cm} \text{parameter map} \ (d, \alpha) \mapsto \beta

\( \psi \) \hspace{1cm} \text{partial inverse map} \ \beta \mapsto (d, \alpha)

\( Q, Q^*, Q_i \) \hspace{1cm} \text{objective functions associated with } \Phi\text{-LASSO}

\( R \) \hspace{1cm} \text{risk function}

\( \mathbb{R}, \mathbb{R}^+ \) \hspace{1cm} \text{set of real numbers, set of non-negative reals}

\( S \) \hspace{1cm} \text{sequence space}

\( \tau, \tau^t, \tau_k^t, L^t \) \hspace{1cm} (t\text{th}) taxon (of the k\text{th} taxon level)

\( \mathcal{T} \) \hspace{1cm} \text{taxonomy}

\( \Theta \) \hspace{1cm} \text{parameter space}

\( V, V_j \) \hspace{1cm} \text{hyper-variable region}

\( X, X_s \) \hspace{1cm} \text{predictor matrix/OTU relative abundances matrix, species matrix}

\( Y \) \hspace{1cm} \text{response}

\( \| \cdot \|_p \) \hspace{1cm} \text{\( l_p \) - norm}

\( \ll, \ll_p \) \hspace{1cm} \text{big ‘oh’, big ‘oh’ in probability}

\( \asymp \) \hspace{1cm} \text{the ratio of two sequences converge to a positive constant}

\( \rightsquigarrow \) \hspace{1cm} \text{convergence in distribution}

\( \wedge \) \hspace{1cm} \text{minimum}
1

Introduction

Metagenomics is the study of the collective genome of a group of organisms in a defined ecosystem, consisting frequently of the study of genetic material recovered from the environment. We contrast this with more traditional genomics, which concerns itself with ‘clonal’ populations, that is, of fairly large genetic homogeneity. We are particularly interested in bacterial populations, which we will refer to as microbiota. While microbiota generally include other micro-organisms such as viruses and protozoa, they will not form part of our data and so we exclude them from our working definition. The microbiome is the corresponding genetic composition of a microbial community.

Microbial systems are ubiquitous to life on earth. They are associated with every organism of higher complexity, as well as separately in soil and water systems. For example, nitrogen fixing Rhizobia in legumes, or Actinobacteria and Frankia in the alder used in environmental remediation, Sprent and Sprent (1990); Lefrançois et al. (2010). In humans they are associated with various microenvironments on the skin and in the gastrointestinal tract; the Human Microbiome Project (HMP) was established to map these locations and
characterize the microbiomes, Turnbaugh et al. (2007). Microbiome dysbiosis is implicated with various health issues, such as antibiotic-associated diarrhea, Loo et al. (2005); Pepin et al. (2005); Poutanen and Simor (2004); Vonberg et al. (2008), and obesity, Bäckhed et al. (2004); Ley et al. (2005, 2006); Turnbaugh et al. (2006).

In the past few decades, our ability to observe systems of microbes has advanced greatly. With the advent of **massive parallel sequencing (MPS)** technologies, we are in a position to capture still frames of diverse microbial communities in time. MPS enables us to study interaction networks between bacteria and correlate microbial states with the progression of disease states in humans, for example. It further allows us to include organisms currently unculturable and heretofore unknown. Every novel data type requires the development of statistical methods particular to its idiosyncracies. Work prior to MPS has required the culturing and typing of the various bacteria under study, an expansive and expensive undertaking.

As applied to bacteria, MPS data frequently consists of the DNA sequences of a hypervariable region \( V \) of a sample of 16S rRNA genes considered representative of a microbiome. No subregion \( V \) is sufficient for discrimination amongst all bacterial species, Chakravorty et al. (2007). Alternative application of MPS is on functional genes, or in shotgun sequencing where the possible metabolic pathways in a community can be mapped out, although these do not account for regulation of gene expression. In a similar guise, transcriptomics considers the DNA that is actually transcribed into RNA, and so allows investigators to probe the active genetic pathways; these methods have progressed from the crude microarray to RNA-sequencing, Chistoserdova (2009); Horgan and Kenny (2011). A wealth of literature has been published using MPS data in metagenomic analyses. Appli-
In order to build predictive models with 16S data, sequences have been clustered into operational taxonomical units (OTUs) to reduce dimension and error. OTUs are produced using a sequence similarity criterion applied as an operational, i.e. data-driven, binning procedure. Once clustered in such a way, microbiomes are treated as vectors of OTU counts. We remark that our definition of OTU is more general than the literature, as we include phylotyping under this umbrella term, Rush et al. (2012). The OTU method is a good first approximation for reducing error, in that it takes into account that the more similar two sequences are, the more likely that one of them was generated in error. The advantage is that it confronts the curse of dimensionality in terms of computing power and singular values. However beyond a threshold, it ignores any further similarity-based error or functional similarity of closely related organisms. Locally (i.e. genetically similarly),
bacteria share similar functionality. Consider the phylogenetic tree in figure 1.1. Intuitively, we mean that for a microbiome with bacteria \( A, B, C, D, \) and \( E \), that \( \alpha A \) will have an effect similar to \((\alpha - \beta)A + \beta B \) on \( \gamma C, \delta D, \) and \( \epsilon E \), where \( 0 \leq \beta \leq \alpha \) and \( 0 \leq \gamma, \delta, \epsilon \).

A phylogenetic model, one that accounts for phylogenetic similarity, allows us to treat random diffusion in a local way. This is important given that the greater similarity in function that a clade exhibits, the less selection between its constituents.

The data is of a high dimensional nature, typically \( p > 1000 \) parameters, and generally paired with low sample size, \( n \ll p \). Classical statistical methods, designed for \( p = o(n) \) situations, no longer apply. We require recourse to model selection methods designed for the \( n \ll p \) regime. In particular, we consider regularized likelihood methods for model selection, of which ridge and LASSO are the most famous. Building on previous work in group model selection, we develop the phylogenetic LASSO (\( \Phi \)-LASSO), a regularization method that incorporates the taxonomic structure induced by the OTUs.

Chapter 2 reviews the goals of modeling in the context of generalized linear models and discusses regularized likelihood methods. Chapter 3 reviews the necessary biology for metagenomics and discusses the definition of bacteria in order to justify use of the 16S rRNA gene in surveying microbial communities. Chapter 4 develops the \( \Phi \)-LASSO and presents results pertaining to its consistency as an estimator. Simulations are performed to demonstrate performance of \( \Phi \)-LASSO. Chapter 5 applies the \( \Phi \)-LASSO to the microbiomes associated with a clinical study using fecal microbiota transplantation (FMT) in the treatment of \( Clostridium \) difficile infection (CDI). Chapter 6 concludes this thesis. Finally, the code for fitting the \( \Phi \)-LASSO is detailed in the appendix.
Model Selection

“Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparametrization is often the mark of mediocrity.”

– Box (1976)

We first introduce the twin model selection and estimation problems before reviewing generalized linear models. We then describe the competition between parameter estimation and model selection and review the model selection literature for regularized regression.
2.1 Models and estimators

Statistical modeling involves a pair of problems: model selection and parameter estimation. We define these using the notion of a statistical model as adopted in McCullagh (2002).

**Definition 2.1.0.1** Let $S$ be the sample space and $\mathcal{P}(S)$ the set of distributions on $S$. A **statistical model** is a map $P : \Theta \rightarrow \mathcal{P}(S)$, where $\Theta$ is a set, called the **parameter space**, parametrizing the distribution. A **submodel** is a restriction $P|_{\Theta'}$ of $P$ to the subspace $\Theta' \subseteq \Theta$.

The **model selection problem** is to select a submodel of the model $P$.

**Definition 2.1.0.2** Consider the model $P : \Theta \rightarrow \mathcal{P}(S)$. An **estimator** $\hat{\theta}$ is a sequence of functions $\{\hat{\theta}_n\}_{n=1}^{\infty}, \hat{\theta}_n : S^n \rightarrow \Theta$. Let $P_\theta \in \mathcal{P}(S)$ and $s_n \in S^n$, where $s_{ni} \sim P_\theta$ for $i = 1, \ldots, n$. Then $\hat{\theta}_n(s_n)$ is called an **estimate**, given the data $s_n$.

The **parameter estimation problem** is to choose $\hat{\theta}$ such that $\hat{\theta}_n$ converges to $\theta$ for almost all $\theta$.

In some contexts, there is a true submodel, a model from which all parameters $\theta$ are found. The estimator that knows this true model is called an oracle. Formally,

**Definition 2.1.0.3** Consider the model $P : \Theta \rightarrow \mathcal{P}(S)$. An **oracle estimator** $\hat{\theta}$ with respect to the subspace $\Theta' \subseteq \Theta$ is an estimator such that $\hat{\theta}_n : S^n \rightarrow \Theta'$.
2.2 Generalized linear models

We orient our work towards generalized linear models (GLMs), introduced by Nelder and Wedderburn (1972) to extend regression methods from the Gaussian linear model to exponential families of distributions and to unify the methods already developed for the disparate models. We source our presentation to McCullagh and Nelder (1989).

**Definition 2.2.1** A random variable $Z$ belongs to an exponential family if its probability density/mass function $f$ takes the form

$$f(z; \theta, \varphi) = \exp \left( \frac{z \theta - b(\theta)}{a(\varphi)} + c(z, \varphi) \right)$$  \hspace{1cm} (2.2.1)

where $a, b, c : \mathbb{R} \to \mathbb{R}$ are known.

Exponential families include the Gaussian, inverse Gaussian, binomial, and many other commonly used distributions. These can be simply extended to multi-dimensional $\theta$.

A generalized linear model consists of three components: (i) a random component given by the exponential family $f$; (ii) a systematic component $\eta = \beta_0 + X \beta$ where $\eta$ is a linear combination of the covariates $X$ parameterized by $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$; and (iii) a link function $g$ connecting the random and systematic components. Altogether, a generalized linear model is given by

$$\mathbb{E}(Y; X) = g^{-1}(\eta)$$ \hspace{1cm} (2.2.2)

where $Y$ is conditioned on $X$ according to $f(y; g^{-1}(\eta), \varphi)$. Point estimates are typically obtained by maximum likelihood estimation. The associated estimation problem is well-defined when the sample size $n$ exceeds the number of parameters $p$. However when $p \geq n$ more complex methods are required. We address this in succeeding sections.
We introduce two important concepts in model estimation: consistency and efficiency. An estimator is consistent if it converges in probability to the true parameter with some specifiable rate, while efficiency is defined relative to the best rate achievable, the $\sqrt{n}$-rate.

**Definition 2.2.2** Let $\hat{\beta}_n$ be an estimator of the true parameter $\beta^0$. Let $h : \mathbb{R} \rightarrow \mathbb{R}^+$ be a positive function with $\lim_{n \to \infty} h(n) = 0$. Then $\hat{\beta}_n$ is said to be a **consistent** estimator of $\beta^0$ with rate $h(n)$ if $||\hat{\beta} - \beta^0|| = \mathcal{O}_p(h(n))$.

The Cramér-Rao lower bound effectively bounds this rate below by $n^{-1/2}$. Formally,

**Theorem 2.2.0.1 (Cramér-Rao Inequality; Garthwaite et al. (2002))** Let $X_1, \ldots, X_n$ be random variables distributed according to $f(x; \beta)$, with the following two regularity conditions:

(i) The Fisher information matrix, $\mathcal{I}(\beta)$, exists and is bounded over the parameter space.

(ii) The operations of differentiation with respect to the parameter $\beta$ and integration with respect to the variable $x$ commute.

Then for any unbiased estimator $\hat{\beta}$ of $\beta$, we have

$$\text{diag}(\mathcal{I}(\beta))^{-1} \leq \text{var}(\hat{\beta})$$

where

$$\mathcal{I}(\beta) = E_{\beta} \left[ \frac{\partial \ell}{\partial \beta^T} \frac{\partial \ell}{\partial \beta} \right]$$

is the Fisher information matrix evaluated at $\beta$, $\ell$ is the log-likelihood, and the inequality is understood entry-wise between vectors.
A stronger version relating to the inverse Fisher information and covariance matrices also exists, see for example Bobrovsky et al. (2013).

Any estimator achieving the lower bound is said to be an efficient estimator. Any estimator beating the lower bound is said to be superefficient. It is a fact that an estimator can beat the Cramér-Rao bound on at most a set of Lebesgue measure zero, Le Cam (1953). A simple example of a superefficient estimator is Hodges’ estimator $\hat{\mu}^H$ of the mean, van der Vaart (1997). This is defined as

$$
\hat{\mu}^H = \begin{cases} 
\bar{x} & \text{for } |\bar{x}| \geq n^{-1/4}, \\
0 & \text{otherwise},
\end{cases}
$$

where $\bar{x}$ is the usual sample mean. $\hat{\mu}^H$ beats the unbiased estimator $\bar{x}$ at a single point in the parameter space, $\mu = 0$. In return, it acquires poor risk properties from non-uniform convergence over the parameter space. We will return to this when discussing model selection consistency in Section 2.3.2.

### 2.3 Competing interests

It is an unfortunate reality that we cannot simultaneously optimize with respect to estimation, prediction, and parsimony. This is especially apparent when the number of parameters $p$ is large. Unbiased estimators can be unstable, even for the Gaussian model, Hoerl and Kennard (1970). Unbiased estimation of a parameter vector $\beta \in \mathbb{R}^p$ with $p$ large often leads to overfitting, where the estimate $\hat{\beta}$ is fitted to the error in the training data and so performs poorly in a predictive setting.
We illustrate this via a simulation of a Gaussian linear model, \( Y_i = X_i \beta + \epsilon_i \) for random response \( Y_i \), covariates \( X_i \in \mathbb{R}^{40} \), parameter vector \( \beta \in \mathbb{R}^{40} \), and random Gaussian error \( \epsilon_i \sim N(0, \sigma^2_i) \). The noise \( \sigma_i \) is taken as alternatively 1 and 20, and the covariance \( \sigma_{jk} \) taken as alternatively 0 and 0.5, \( j \neq k \), and \( \sigma_{jj} = 1 \). Here we compare three methods of fitting the model to the oracle estimator, the estimation made when the true model is known. The three methods are ordinary least squares (OLS), ridge, and LASSO, these last two defined in Section 2.4. The intuition is that LASSO selects covariates with high signal strength and ridge reduces variation in estimation. OLS provides unbiased estimation of \( \beta \). However, it is trumped by LASSO and ridge in terms of predictive performance. While the LASSO estimate \( \hat{\beta}_{\text{LASSO}} \) is more easily interpreted, it does not uniformly dominate the ridge estimate \( \hat{\beta}_{\text{ridge}} \). The results for estimation and prediction error, defined as \( \text{SSE} = \sum_{j=1}^{p}(\beta_j - \hat{\beta}_j)^2 \) and \( \text{SSPE} = \sum_{i=1}^{n}(y - x\hat{\beta})^2 \), are presented in Tables 2.1 and 2.2, respectively. Note that even the oracle estimator struggles with overfitting when the true parameter space is high-dimensional and is outperformed by ridge and LASSO, see the prediction error.

The lack of dominance of ridge and LASSO is illustrative of another issue, (sometimes) spurious correlation. The LASSO is minimax optimal for certain conditions on the design matrix \( X \), Ye and Zhang (2010); Raskutti et al. (2011). However, when covariates exhibit group-wise correlation, it tends to favour selection of the most uncorrelated covariates, for which the ridge method is less susceptible, Tibshirani (1996); Zou and Hastie (2005). Correlation and multicollinearity are especially problematic in the \( p \gg n \) regime, where model identifiability is an issue since we can fit the data perfectly in multiple ways. It is necessary to balance estimation bias with model identifiability.
Table 2.1: Median estimation error (SSE) for Gaussian model $Y = X\beta + \epsilon$ where $X \in \mathbb{R}^{50 \times 40}$ with entries drawn from the standard normal, with $\epsilon \sim N(0, \sigma_\epsilon)$, based on 100 replicates. The interquartile range is provided in parentheses. Simulation described in Section 2.3.

<table>
<thead>
<tr>
<th>Covariance ($\sigma_{jk}$)</th>
<th>Noise-to-signal ($\sigma_\epsilon$)</th>
<th>1</th>
<th>$1/\sqrt{2}$</th>
<th>0</th>
<th>$1/\sqrt{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>0.710(0.425)</td>
<td>1.29(0.62)</td>
<td>270(145)</td>
<td>537(282)</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>4.28(3.31)</td>
<td>8.46(4.99)</td>
<td>1770(1150)</td>
<td>3110(2330)</td>
<td></td>
</tr>
<tr>
<td>Ridge</td>
<td>2.66(1.30)</td>
<td>3.07(1.43)</td>
<td>24.3(15.2)</td>
<td>13.4(10.4)</td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>2.26(1.41)</td>
<td>2.51(1.79)</td>
<td>29.7(23.8)</td>
<td>88.6(32.2)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Median prediction error (SSPE) for Gaussian model $Y = X\beta + \epsilon$ where $X \in \mathbb{R}^{50 \times 40}$ with entries drawn from the standard normal, with $\epsilon \sim N(0, \sigma_\epsilon)$, based on 100 replicates. The interquartile range is provided in parentheses. Simulation described in Section 2.3.

<table>
<thead>
<tr>
<th>Covariance ($\sigma_{jk}$)</th>
<th>Noise-to-signal ($\sigma_\epsilon$)</th>
<th>1</th>
<th>$1/\sqrt{2}$</th>
<th>0</th>
<th>$1/\sqrt{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>2030(530)</td>
<td>188(496)</td>
<td>29000(7300)</td>
<td>71800(194800)</td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2050(530)</td>
<td>188(494)</td>
<td>36600(9200)</td>
<td>72400(193400)</td>
<td></td>
</tr>
<tr>
<td>Ridge</td>
<td>181(80)</td>
<td>1.12(3.11)</td>
<td>20800(5200)</td>
<td>190(512)</td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>160(69)</td>
<td>1.07(2.99)</td>
<td>21200(5300)</td>
<td>206(558)</td>
<td></td>
</tr>
</tbody>
</table>

2.3.1 Estimation and Risk

We can quantify our success in achieving a global objective in terms of risk. Every decision carries consequences, which we attempt to capture in terms of a loss function. Risk is simply the expected loss from making a decision marginalized over the sampling space $\chi$. We source our treatment from Garthwaite et al. (2002); Berger (1985). The interplay of estimator risk and efficiency become apparent when we consider consistency in model selection in Section 2.3.2.

Definition 2.3.1 Let $F(x; \beta)$ be the distribution function of random variables $X \in \chi$, conditioned on parameters $\beta \in \Theta$, and let $L(\beta, \hat{\beta})$ be the loss from estimating $\beta$ by $\hat{\beta}$. The
risk function $R$ is defined as

$$R(\beta, \hat{\beta}) = \int_X L(\beta, \hat{\beta}) dF(x; \beta).$$

Two important classes of estimators minimize the maximal and expected risk, respectively.

**Definition 2.3.2** An estimator $\hat{\beta}$ minimizing the maximal risk $\sup_{\beta} R(\beta, \hat{\beta})$ is said to be minimax optimal. An estimator $\tilde{\beta}$ minimizing the Bayes risk $E_{\tilde{\beta}}(R) = \int_{\Theta} R(\beta, \tilde{\beta}) \pi(\beta) d\beta$ with respect to a prior distribution $\pi(\beta)$ is said to be Bayes optimal.

Maximal risk considers exclusively the worst case scenario, while Bayes risk weights all scenarios according to their purported prior probability.

### 2.3.2 Model Selection Consistency

We seek to characterize how well we select the correct model. Do we achieve consistent model selection, and does this agree asymptotically with the oracle estimator?

**Definition 2.3.3** Let $\beta_n^0$ be the true parameter vector, and let $A_n^0 = \{j : \beta_{nj}^0 \neq 0\}$. Let $\hat{\beta}_n$ be an estimator of $\beta_n^0$ and let $A_n = \{j : \hat{\beta}_{nj} \neq 0\}$. We say that $\hat{\beta}_n$ is a sparse-type estimator of $\beta_n^0$ if $P\{A_n \subseteq A_n^0\} \to 1$ as $n \to \infty$. More strongly, we say that $\hat{\beta}_n$ is model consistent if $P\{A_n = A_n^0\} \to 1$ as $n \to \infty$.

These properties relate to the convergence of the estimator to the true model, or a submodel thereof. Note both estimator and parameter vectors are subscripted by $n$, to accommodate the growth of parameter dimension $p_n$ as $n$ increases. We are also interested in the efficiency of estimation. We adopt the following form of asymptotic normality, as considered in Fan and Peng (2004); Zhou and Zhu (2010).
**Definition 2.3.4** Let $\hat{\beta}_n$ be a model consistent estimator of $\beta_0^n$. Suppose $\hat{\beta}_n$ satisfies the $\sqrt{n}$-law

\[ \sqrt{n} A_n I_n^{1/2} (\hat{\beta}_n^0 A_n - \beta_0^n A_n) \to N(0, \Sigma) \]

where $A_n$ is a $q \times |A_n|$ matrix such that $A_n A_n^T \to \Sigma$, $\Sigma$ is positive semidefinite, and $I_n(\beta_0^n)$ is the Fisher information matrix evaluated at $\beta_0^n$. Then $\hat{\beta}_n$ is said to have the **oracle property**.

The asymptotic normality described in Definition 2.3.4 says that the estimator performs asymptotically as well as the oracle estimator.

Although superficially an attractive property, sparsity can be problematic. Sparse-type estimators are examples of superefficient estimators, described in Section 2.2. Leeb and Pötscher point out these are similar to Hodges’ estimator of the mean, with similar risk properties, Leeb and Pötscher (2008). Considering loss functions of the form $l(\hat{\beta} - \beta)$, they show that if an estimator is sparse-type, then its maximal risk is equal to the supremum of the loss function, irrespective of the loss. It is in fact more insidious, for through a simple modification of their argument, we can show that if $\hat{\beta}_n$ is sometimes of sparse-type, i.e. $\mathbb{P}\{A_n \subseteq A_n^0\} \to c > 0$, and the loss function is unbounded, then $\hat{\beta}_n$ has worst maximal risk. Like Hodges’ estimator, the problem is non-uniformity of convergence over the parameter space $\beta$. This is a stern reminder that the asymptotics of an estimator are of little merit if they perform poorly in finite samples, a regime to which we are physically restricted, Pötscher (1991).

Leeb and Pötscher consider only maximal risk, but other risks exist, reflecting other priorities. We introduce sparsity due to a prior belief that solutions are sparse
or because we seek a sparse, parsimonious, solution. From a Bayesian perspective, we incorporate a sparse prior on the parameter space. In this situation maximal risk is too coarse a measure. It makes sense instead to evaluate such an estimator in terms of Bayes risk, the expected risk given the prior.

### 2.4 Regularized Maximum Likelihood

Regularized maximum likelihood estimation incorporates a penalty term in the estimation of model parameters to improve a variety of properties, including prediction error, estimator variance, model complexity, and algorithmic stability. Formally, we have

**Definition 2.4.1** Consider the loss function

\[ Q(\beta) = \ell(\beta; Y, X) - p_\lambda(\beta) \]  \hspace{1cm} (2.4.3)

where \( \ell \) is the log-likelihood corresponding to model (2.2.2) and \( p_\lambda : \mathbb{R}^p \to \mathbb{R} \) is a class of penalty functions indexed by hyper-parameter vector \( \lambda \). Then the regularized maximum likelihood estimate (r-MLE) \( \hat{\beta} \) is given by

\[ \arg\max_\beta Q(\beta) \]  \hspace{1cm} (2.4.4)

Note that we do not penalize the intercept term \( \beta_0 \). We have not explicitly included \( \beta_0 \) to reduce the visual complexity. We further remark that, although we treat GLMs, the likelihood \( \ell \) is not restricted to this setting, nor need we only consider likelihood functions.

Two of the simplest r-MLEs employ \( \ell_1 \)- and \( \ell_2 \)-norms in their penalty.

**Definition 2.4.2** Let \( \lambda \in \mathbb{R}^+ \).
(i) If $p_\lambda(\beta) = \lambda ||\beta||_2^2$, then $\hat{\beta}$ is the **ridge** estimator;

(ii) If $p_\lambda(\beta) = \lambda ||\beta||_1$, then $\hat{\beta}$ is the **least absolute shrinkage and selection operator** (LASSO) estimator.

Use of the ridge precedes the LASSO. It was introduced for Gaussian linear models to control inflation and instability of the MLE solution when the design matrix $X$ has one or several nearly singular values, Hoerl (1962); Hoerl and Kennard (1970). This is encapsulated by the relation $E[(\beta^0 - \hat{\beta}_{ols})'(\beta^0 - \hat{\beta}_{ols})] = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}$ where $\hat{\beta}_{ols}$ is the ordinary least squares estimate of $\beta^0$, $\sigma^2$ the common variance, and $\{\lambda_i\}_{i=1}^{p}$ are the eigenvalues of $X^tX$, with corresponding variance $\text{var}[(\beta^0 - \hat{\beta}_{ols})'(\beta^0 - \hat{\beta}_{ols})] = 2\sigma^4 \sum_{i=1}^{p} \lambda_i^{-2}$. Not just of theoretical importance, this is important in finite arithmetic; the `lm` method in R employs a small ridge penalty when $X^tX$ is nearly singular, R Core Team (2016).

Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) to linear regression in order to perform simultaneous estimation and variable selection. It is able to estimate exactly zero due to its non-differentiability at the boundaries of the orthants. Similar to ridge regression, it exhibits reduced inflation and instability relative to the OLS estimator in nearly singular situations. Significance tests are introduced in Lockhart et al. (2014). Zou (2006) proposes the adaptive LASSO, which possesses the oracle property under certain conditions. A number of oracle results pertaining to the LASSO are collected in van de Geer and Bühlmann (2009). Under certain conditions on the tuning parameter, the (adaptive) LASSO is (nearly) minimax optimal. In light of the discussion in Section 2.3.2, they can not be both model consistent and minimax optimal.

Frank and Friedman (1993) studied the continuum of **bridge** estimators, with
\[ p_\lambda(\beta) = \lambda \sum_{j=1}^{p} | \beta_j |^\gamma \] for fixed \( \gamma \in (0, 1] \), in application to chemometrics. These may be extended to \( \gamma \in \mathbb{R}^+ \) and include the ridge and LASSO, at \( \gamma = 2 \) and \( \gamma = 1 \), respectively. Fu and Knight (2000) consider the asymptotics of these estimators, which they call LASSO-type. These have found favour in compressive sensing for \( \gamma < 1 \), Chartrand (2007).

When the covariates of the design matrix \( X \) exhibit high correlations, the LASSO tends to select the least correlated predictors, Tibshirani (1996). On the other hand, ridge regression has been observed to shrink correlated predictors as a group but performs no selection. Zou and Hastie (2005) proposed the \textbf{elastic net}, with \( p_\lambda(\beta) = \lambda_1 | | \beta | |_1 + \lambda_2 | | \beta | |_2^2 \) as a compromise. It performs variable selection, and due to the grouping behaviour of the \( l_2 \)-norm, correlated groups of variables tend to be in or out of the model together. Similar to the LASSO, there is an adaptive version of the elastic enjoying oracle properties, Zou and Zhang (2009).

2.4.1 Convexity

The bridge estimators with \( \gamma \geq 1 \) and the elastic net are examples of convex estimators, due to the convexity of their associated penalties.

**Definition 2.4.1.1** An estimator is called \textbf{convex} if the associated optimization problem is convex. Otherwise the estimator is \textbf{non-convex}.

For example, the ordinary least squares estimator for a Gaussian model is convex. This convexity has permitted the design and implementation of several efficient fitting algorithms. For instance least angle regression (LARS), Efron et al. (2004); Zou and Hastie (2005); Yuan and Zou (2009), local linear approximation (LLA), Zou and Li (2008), cyclic
coordinate descent, Wu and Lange (2008); Friedman et al. (2010), and the concave convex procedure (CCCP), Kim et al. (2008); Wang et al. (2013). These are often improved with path-following implementations at low additional computational cost. Here path-following refers to the use of previous solutions at the last tuning value $\lambda$ as warm starts for the next. There are also many convex estimators incorporating grouping information, which we describe in Section 2.4.2.

Non-convex penalties have also been proposed, Fan and Li (2001); Fan and Peng (2004). These include bridge estimators for $\gamma \leq 1$. Other non-convex penalties include the smoothly clipped absolute deviation (SCAD), Fan and Li (2001), minimax concave penalty (MCP), Zhang (2010a), and more generally folded-concave penalties, Lv and Fan (2009); Fan and Lv (2011).

**Definition 2.4.1.2 (SCAD)** The smoothly clipped absolute deviation penalty $p^{SCAD}_\lambda(z)$ is defined

$$
\begin{align*}
|z| & \quad \text{if } |z| \leq \lambda \\
-\frac{|z|^2 - 2a\lambda|z| + \lambda^2}{2(a+1)} & \quad \text{if } \lambda < |z| \leq a\lambda \\
\frac{(a+1)\lambda^2}{2} & \quad \text{otherwise}
\end{align*}
$$

where $a, \lambda$ are positive constants and $z \in \mathbb{R}$

**Definition 2.4.1.3 (MCP)** The minimax concave penalty $p^{MCP}_\lambda(z)$ is defined

$$
\lambda \int_0^z (1 - x/(\gamma \lambda))_+ dx
$$

where $\gamma, \lambda$ are positive constants, $z \in \mathbb{R}$, and $(x)_+ := \max(0, x)$. 
The advantage of these non-convex estimators lies in their reduced bias relative to convex penalties, Fan and Li (2001). Further, the sparsity of the solutions often give rise to the oracle property. It is a simple consequence of the group oracle result in Zhou and Zhu (2010) that the oracle property holds for the $l_{1/2}$-penalty. In fact, we will see in Chapter 4 that the oracle property holds for the $l_{1/m}$-penalty, for all natural numbers $m \in \mathbb{N}, m \geq 2$.

The main disadvantage in non-convex regularization is in solving the corresponding optimization program. Non-convexity gives rise to local optima that are not globally optimal. Recent investigations have considered the entire solution path in the demonstration of estimation and model selection consistency, Wang et al. (2013); Yukawa and Amari (2016). Fan and Lv have shown that particular care must be made to ensure the solution path for the fitting algorithm contains the oracle estimator with asymptotic probability 1, Fan and Lv (2011).

While computationally difficult, work has progressed on fitting non-convex r-MLEs. Breheny and Juang (2011) develop coordinate descent algorithms for non-convex r-MLEs, including MCP and SCAD, an improvement over the previously proposed local linear approximations. Wang et al. (2014) show that the CCCP produces consistent estimates whose solutions paths contain the oracle estimator with asymptotic probability 1. Initializing the gradient descent or LLA algorithm from a LASSO solution attains local solutions with the oracle property Zhang (2010b); Zhang and Zhang (2012); Fan et al. (2014).
2.4.2 Structured sparsity and group selection

For the r-MLEs considered so far, we have penalized each parameter separately. However, there are situations in which groups of covariates should be in or out of a model together (group selection) or where one covariate should be included before another (hierarchy). Regularized group selection has been studied as early as 1999 with the group-LASSO, a simple generalization of LASSO which shrinks and selects \textit{a priori} defined groups, Bakin (1999).

The group-LASSO employs the penalty \( \sum_{k=1}^{K} ||\beta_k||_2 \), where \( \beta_k \in \mathbb{R}^{pk} \) is the sub-parameter vector corresponding to group \( k \). The absence of a square on the \( l^2 \)-norm leads to non-differentiability at \( \beta_k = 0 \), resulting in sticky estimation at zero. Extensions of the group-LASSO to the GLM framework with efficient algorithms are considered in Meier et al. (2008); Roth and Fischer (2008).

Polynomial models should satisfy strong hierarchy, where lower order terms need to be included before higher order terms, since a translation in the base covariates almost surely results in the appearance of intermediate monomials, consider for example the polynomial \( (x_1 - \alpha)^3(x_2 - \beta)^3 \). Bien et al. (2013) develop LASSO-like methods satisfying strong hierarchy. Lim and Hastie (2015) develop a regularization method which does not directly penalize main and interaction effects. Zhao et al. (2009) develop the CAP estimators for grouped and hierarchical models, issuing from a combination of group bridge-like penalties all with \( \gamma \geq 1 \).

Yuan and Lin (2006) emphasize the importance of group-wise orthogonal invariance for categorical variables where the factor has multiple levels and hence non-unique
representation by indicator functions. This refers to invariance of estimation under orthogonal transformations within-group. In this way the individual components of a factor are included or excluded together, and the estimates do not depend on representation of the factor. Another situation potentially requiring orthogonal invariance is when spatial distribution is used in prediction, where regularized estimates could otherwise depend on the coordinate system. The group-LASSO itself is within-group orthogonally invariant.

A non-convex version of the group-LASSO is developed by Zhou and Zhu (2010), which the authors call the hierarchical lasso (HLASSO). This penalizes the root of the $l_1$-norm of each group, $\sqrt{||\beta_k||_1}$. A simple consideration of the Pythagorean Identity reveals this cannot be group-wise orthogonally invariant. We will consider the HLASSO in more detail in Chapter 4, where we extend it to arbitrary grouping levels.

Jenatton et al. (2011) explore sparsity patterns for structured variable selection, generalizing a number of convex methods in current use, such as the CAP estimators. They consider exclusively convex penalties.

In recent years, attention has turned towards tree-guided selection methods, where covariates may be grouped according to some tree or directed acyclic graphs, for example Liu and Ye (2010). Kim and Xing (2012); Garcia et al. (2013); Zhang et al. (2015) explicitly consider biological applications. All use convex penalties, and none consider phylogeny. We will pursue this idea further in Chapter 4.
21

3

Metagenomics

“[T]he abiding intellectual scandal of bacteriology has been the absence of a clear concept
of bacterium”

– Stanier and van Niel (1962)

These words are related in the introduction to The Prokaryotes, 4e, Woese (2013). They were made in a nascent age of molecular probing and characterization of organisms. Indeed, the first full genome to be sequenced was for Bacteriophage MS2, Fiers et al. (1976), and that of the first bacterial organism was Haemophilus influenzae, Fleischmann et al. (1995). Only a few years following their despairing remarks, the idea of using genes as molecular clocks was introduced, Zuckerbrandt and Pauling (1965). This was followed by the rRNA delineation of the three domain tree of life, Woese and Fox (1977); Woese (1987).

We are interested in the composition of microbial systems, i.e. the microbiome, whose characterization we pursue through 16S rRNA amplicon metagenomic analysis. In this chapter, we describe metagenomics. In particular, we describe the data central to
this thesis, used to characterize the bacterial composition of microbial communities. We consider further some of the ontological issues in defining bacteria, inherited from their origins to their modes of propagation, introducing a notion of bacteria in aggregate which unifies efforts to define bacteria and their sub-types.

3.1 Information System

We consider three types of biological sequences and their relationships: DNA, RNA, and peptides. DNA and RNA form sequences over the alphabets \( \{A,C,G,T\} \) and \( \{A,C,G,U\} \), respectively, and encode the same information. Peptides are sequences over a larger alphabet of amino acids. DNA readily assembles into double-stranded helices, and consequently forms a stable information repository, while RNA and peptides have acquired generally more functional roles in bacterial development. Together these form a tightly coupled core information processing system along with more loosely associated peripheral functions.

The classical view, christened “The Central Dogma”, held that the flow of information was unidirectional, with RNA transcribed from DNA and subsequently translated into peptides, which form proteins, Crick (1970). The transcription from DNA to RNA is an isomorphism induced by the bijective mapping between the DNA and RNA bases. A three base sequence, called a codon, codes for a single amino acid. While over 500 amino acids are known, Wagner and Musso (1983), only 20 amino acids are encoded by a codon, so that the mapping RNA→peptide is non-injective. For this reason DNA is commonly called a degenerate code. This redundancy allows some variation in the genetic code without
altering the corresponding protein.

Investigations in the intervening years have revealed a more nuanced process. The two leading origin of cellular life theories give priority to RNA over DNA, with DNA gradually acquiring the role of information reservoir from RNA, Wächtershäuser (2013). Indeed, transcription and translation are performed by proteins and protein-RNA couplings, respectively. The complexity of this self-referential information system ensures that few changes occur to it that do not adversely affect the organism hosting it and its progeny. Other, more peripheral cell functions may be modified or disconnected with greater flexibility, for example genes encoding antibiotic resistance, Davies and Davies (2010). The preservation of this central system has encouraged its use as the basis for a prokaryotic phylogeny, Woese (2013).

The double-stranded nature of genomic DNA is induced by the complementarity of the bases, with A associating with T and G associating with C. We describe this formally as follows.

**Definition 3.1.0.1** A DNA sequence is an element of the free product $S = \cup_k \{A, C, G, T\}$.

DNA sequences have an orientation determined by the linkage of the nucleic acids on the sugar backbone. We denote this orientation $5'-s-3'$. For example, $5'-\text{AGT}-3'$ is the same sequence as $3'-\text{TGA}-5'$ and distinguished from $5'-\text{TGA}-3'$.

**Definition 3.1.0.2** The pairs A and T, and C and G form complementary bases. The complement of a DNA sequence is obtained by replacing the bases with their complementary bases and then reversing orientation.
Thus 5′ − ACT − 3′ is the complement of 5′ − AGT − 3′. Double-stranded DNA appears as two complementary sequences, for example

\[
5' - AGTCG - 3' \\
3' - TCAGC - 5' 
\]

DNA sequences sometimes share common origins. This is captured using the notion of sequence homology. Since homology is difficult to describe formally, we will consider a restricted definition excluding items such as pseudogenes and change in function. Our notion is based on functional continuity. Consider a DNA sequence encoding some function. There exists a subspace \( E \) of the free product \( S \) consisting of those sequences encoding the same function. A change in a sequence is called a mutation. There are two main classes of these. **Point mutations** occur when one base is substituted for another in the sequence, e.g.

\[
5' - AGTCG - 3' \Rightarrow 5' - AGCCG - 3' 
\]

**Insertions and deletions** occur when a contiguous subsequence is inserted or excised from the original sequence, respectively, e.g.

\[
5' - AGTCG - 3' \Rightarrow 5' - AGT\text{GCG} - 3' \quad \text{(insertion)} 
\]

Two sequences are ancestrally related, or **homologous**, if there exists a plausible mutation path connecting the two. Plausible is here intentionally vague, as it must be informed by biology we will not delve into. However, to provide a concrete example, consider the
Figure 3.1: (a) Distribution of DNA sequences over sequence space $S$. The dotted circles surround clusters representing four separate lineages. (b) Underlying phylogeny of distribution presented in (a).

Following three sequences,

$$5' - \text{GGGGGGAGTCGATACCCCCC} - 3' \quad (3.1.1)$$

$$5' - \text{GCGGGGGAGTCGATACCCCAG} - 3' \quad (3.1.2)$$

$$5' - \text{AAAAAAAGCGAAGACTTTTTTT} - 3' \quad (3.1.3)$$

The obvious two-dimensional structure of the corresponding RNA is folding into loops, with the complementary tails (blue) forming the stem. Sequences (1.1) and (1.2) differ at two locations, any of which could have occurred without ruining the structure. To mutate to (1.3), the entire ‘stem’ needs to be replaced, which is implausible. Thus we would say sequences (1.1) and (1.2) are homologous, while sequence (1.3) is not homologous to either. Thus the subspace $E \subseteq S$ consists of a class of homologous sequences. $E$ is a concept central to our analysis.

A final important concept in the information system under consideration is recombination. Recombination occurs when a contiguous subsequence of a strand of DNA is replaced by another. This is typically preceded by the acquisition of genetic material from
external sources, termed **horizontal gene transfer**. Horizontal gene transfer manifests as one of (i) bacterial conjugation (whereby bacteria share genetic material); (ii) transformation (environmental uptake), or (iii) transduction (bacteriophage-mediated transfer).

### 3.2 A Bacterial Discriminant

We use the 16S rRNA gene to identify bacteria in microbial systems. The 16S gene encodes an RNA subunit of the bacterial ribosome, which is essential to protein production and hence cell function. Further, selective pressure imposed by its function ensures that few mutations occur that do not alter the structure of this rRNA and so adversely affect the viability of the bacterium and its progeny. Its secondary structure relies on conserved, approximately complementary subsequences, introducing correlation between various subsequences, and so redundancy is naturally encoded. Given the time period over which bacteria have existed, a number of mutations have accumulated to permit discrimination among groups. Alternatives to the 16S rRNA gene include the other rRNA genes, functional genes such as *rpoB*, Case et al. (2007), and more generally environmental gene tags, Tringe and Rubin (2005). In view of horizontal gene transfer, it is unlikely any one gene can perfectly recall the species. However, Bowman et al. (2015) remain optimistic, suggesting that if the 16S+23S region could be contiguously sequenced, then the 16S and 23S regions could be used to identify and validate genus, while the internal transcribed spacer separating the two regions could be used for species identification.

The 16S gene currently enjoys greatest favour as a global marker, although this is in part self-fuelling, as it is also the most heavily analyzed marker. It forms the basis
for current classification of bacteria, with sequence similarity of $> 97\%$ typically taken as indicative of common species, Woese (2013). The 16S gene has nine subsequences $V_j$, $j = 1, ..., 9$, which are described as hyper-variable, flanked by regions of high conservation. This means that across bacteria, each collection of homologous sequences $V_j$ exhibit greater variation than the local average when extending the window into the flanking conservative regions; they are still highly conserved. The various $V_j$ have different discriminatory power, providing varying resolution into each bacterial taxon, Chakravorty et al. (2007). Further, sequences can vary widely within species.

The so-called next-generation sequencing (NGS) technologies perform sequencing by synthesis, whereby the DNA sequences are replicated, and the order of the inclusion of new nucleic acids is tracked. This concept has revolutionized sequencing, permitting the simultaneous sequencing of millions or billions of sequences on a single chip. Prior to this, the Sanger method, which adopts the chain termination principle, Sanger et al. (1977), was the dominant method of DNA sequencing. While time-consuming, it produces highly accurate sequences. As recently as 2008 it was the most common method of sequencing, Kunin et al. (2008). Current NGS technologies do not permit full length 16S sequencing on massive scales. The 16S gene is approximately 1500 base pairs (bp) in length. Illumina’s dual-plex sequencing is capable of sequencing a region up to 500 bp, while Roche 454’s pyrosequencing is now allegedly capable of sequencing a 700 bp region. As a compromise, a homologous region comprising one or several contiguous hyper-variable regions are chosen and sequenced, sacrificing resolution for coverage. The general outline is as follows. First, the DNA is extracted from a sample of bacteria. The region $V$ of interest is amplified via the polymerase chain reaction (PCR), using DNA primers designed to associate with
the conserved regions flanking the region $V$. These primers may be fitted with an adap-
tor sequence which encodes the origin of the sample. The $V$ amplicons are submitted to
sequencing, where the sequences of millions or billions of distinct DNA strands are deter-
mined. One or several contiguous hyper-variable regions $V$ form our evolutionary space $E$.

3.2.1 Species proxies

Since the 16S gene varies within species, the sequences obtained do not necessarily
uniquely represent a species. To form approximate species representatives, sequences are
clustered according to one of an assortment of clustering algorithms, according to some
sequence similarity criterion. Additionally, due to PCR and sequencing error, not all se-
quences represent a true species, but rather exist outside the space $E$ of viable sequences.
The resulting clusters are called operational taxonomic units (OTUs), reflecting their
data-driven origin. Ideally these consist of unique modes with low dispersion about them.

Due to variation in clustering methods, the OTUs will vary from method to
method. This presents a problem in using these as predictors, due to diffusion between
OTUs by clustering algorithms, which may merge different modes or draw borders between
OTUs in different ways. We illustrate this in Figures 3.2 and 3.3, where we use $\mathbb{R}^2$ for
analogy. Recall that our data forms a frequency distribution on the free product over DNA
bases, $\bigcup_{k>0}\{A, C, G, T\}^k$, falling on a neighbourhood of the evolutionary space $E$. Using two
alternative clustering algorithms and three different metrics, we obtain different hierarchical
clusterings. In most methods, clusters ‘o’ and ‘x’ achieve perfect separation, while clusters
‘c’ and ‘s’ are always to some degree imperfectly separated. Importantly, the methods all
differ in their success at separating the point clusters.

In addition to clustering into OTUs, sequences are also classified according to a training taxonomy to provide a nearest identification. A (bacterial) taxonomy is a systematic classification of bacterial groups. Broadly, this consists of classifying a well-defined bacterial unit at various levels of evolutionary relation. Encoded in this hierarchy is increasing genetic similarity of the core organism and specific functional properties. In our context, this consists of six taxon levels, the phylum, class, order, family, genus, and species taxon levels. This is described in greater detail in Chapter 4. Since our data consists of subsequences of the 16S gene sampled with error, we cannot perform species classification; even genus classification is dubious in some situations.

It is these notions of random diffusion between OTUs due to variation in clustering method and functional proximity due to phylogenetic proximity that we wish to incorporate into our modelling. For our application, the end product from sequencing and processing is a matrix $X$ of OTU relative abundances as a proxy to the matrix $X_s$ of species relative abundances.

### 3.3 Defining Bacteria

#### 3.3.1 Diversity in Propagation

We consider two modes of propagation with respect to bacteria. The first is cell propagation, the second gene(tic) propagation. The naive intuition, informed by macroscopic observation and analogy to plants and animals, is that these parallel each other. However, the genome of bacterial cells is not entirely stable over its lifetime, which begs
the question as to what whether we should follow genes or cells.

The first mechanism of bacterial reproduction a student learns is binary fission, by which a mother cell divides into two daughter cells. Other methods are multiple fission, budding, and mycelial extension and fragmentation. There is no reason to believe this list complete. In seeming opposition is cell fusion, which can take the form of the phospholipid bilayer of two cells fusing, one cell penetrating the bilayer of another, or one cell engulfing another but failing to digest it. Zinder and Dworkin (2013)

When DNA is transmitted down a cell lineage, this is referred to as vertical gene transfer, and is akin to the transfer of genetic information from parent to child. When a cell obtains genetic information from an external source, it is called horizontal gene transfer (HGT). Perhaps more than any other phenomenon, it is HGT which complicates our notion of ancestry and has historically inhibited the systematic study of bacteria and their phylogeny.
Figure 3.3: Depiction of three different hierarchical clustering methods applied to the points in Figure 3.2. Employed are the complete, average, and single linkage methods with Euclidean, maximum, and Manhattan metrics. See Section 3.2.1.
In light of HGT it is natural to wonder whether there is a limit to which genes or genetic information may be transferred. There are some investigators who believe in some core set of genes which are never transferred horizontally, with genes encoding the central information processing machinery counted among them, Ciccarelli et al. (2006). The justification being that since the functions are so tightly dependent on each other, a large portion of the information needs to be transmitted for viability. However, genes encoding the information system have been reported to undergo HGT events between apparently disparate lineages, with some recombination between closely related lineages, Bansal and Meyer (2002); Del Casale et al. (2011); Harth et al. (2007). There is explicit evidence of 16S rRNA transfer, Yap et al. (1999); Harrington et al. (2012).

3.3.2 Distributional bacteria

We have seen that a degeneracy exists in bacterial propagation. The phenomena of HGT and CF prevent an aggregate view of bacteria as a group of individual agents, while HGT and mutations serve to blur distinctions. These random dynamics suggest an explicitly stochastic definition of bacteria. Rather than seek exact group delineation, track the modes as bifurcation events occur. The divergence of modes can characterize speciation, for example. The advantage of a distributional concept is that it forces us to view speciation as a gradual and perhaps locally reversible process, rather than the temptingly simple one-off event in the past.

Rather than rely on rigid features for explicit definition, we appeal to a feature’s association with others to establish bacterial type. In vitro, this involves various metabolic and genetic assays. The golden standard for delimiting species is via DNA-DNA hybridiza-
tion. In this procedure, we begin with two pools of DNA from separate monocultures. The two pools are combined, and the denatured DNA allowed to associate with approximately complementary strands. The higher the degree of re-association, the more dissimilar the pools are. It is standard to take a hybridization value of less than 70% as indicative of separate species, Wayne et al. (1987). This re-association value is necessarily arrived at empirically. The essential observation is that more dissimilar bacterial groups have a higher joint re-association value. This allows us to speak of a similarity continuum. In fact, the use of a hard re-association value is a symptom of our compulsion to discretize continuous objects so that we may understand them, which has also played a role in the adoption of the Neyman-Pearsonian dogma.

While we have not seen it anywhere made explicit, microbiology wants a distributional definition for bacteria. These are best understood as distributions or stochastic processes of their metagenome. The genetic re-association assay is essentially a laboratory test for statistical distinguishability of distributions. Bacteria are thus viewed, in aggregate, as distributions of genes. In the abstract, bacterial groups are covariance matrices of their metagenome together with their modes. We might define a species or strain as unimodal with low dispersion. A speciation event is recorded by the bifurcation of distributions. Broader groups can then be defined in terms of relaxations of these two conditions. Current definitions of species and lineages rely on choosing marginals providing some degree of practical discrimination and perhaps chosen to match their authors concept of taxon, Kämpfer and Glaeser (2013). Consensus phylogenies from multiple phylogenetic trees attempt to locate the true (non-marginal) modes as they bifurcate in the past. The choice of gene(s) for constructing a phylogeny is problematic. We have focused on the 16S rRNA
gene, however others exist. Thus when we consider bacterial types, we consider the distribution of the metagenome, conditioned on a collection or distribution of core genetic information.

### 3.3.3 Surveying Bacterial Communities

In light of the previous discussion, bacterial communities can be viewed as mixtures of distributions on the metagenome. Surveying microbial systems require a method efficient in time and resources. Ideally a molecular marker exists so that we can distinguish between the various modes of bacterial type. The search for such a marker amounts to a search among the marginals for the clearest modes with highest variation between (non-marginal) modes. We have several desiderata for a marker, among them

(i) the marker is present in all bacteria;

(ii) the marker is highly conserved;

(iii) the marker is varied enough for discrimination among bacterial groups of interest;

(iv) bacteria do not exchange or otherwise acquire new markers;

(v) each bacterium has exactly one such marker.

Selecting from among the genes encoding the bacterial information system would satisfy (i) and (ii). It is these genes which are used to define bacteria (in opposition to archaea and eukarya), and their central importance to cellular function limits the rate of alteration to the system. (iii) is more challenging, although for technical reasons we can satisfy this by construction. (iv) ensures we are not conflating disparate groups, and (v) ensures that
bacteria are not double-counted when the marker is used to survey a microbial system. It is unfortunate that no well-defined marker has been observed satisfying this. In fact, desiderata (v) may be in opposition with (ii), since any highly conserved feature is likely critical to bacterial viability, so that it behooves a bacterium to keep it in multiples.

In light of HGT, the identification of a marker satisfying (iv) is improbable. Even genes encoding the basic machinery of bacteria have been observed to experience HGT events. We are left seeking a marker that is global, redundant, conserved, and hyper-variable. As previously described, a gene satisfying (i)-(iii) exists, the 16S rRNA gene.
The phylogenetic LASSO

“The purpose of models is not to fit the data but to sharpen the question.”


Metagenomics, like the other ‘-omics’, bears a structural relationship between covariates. Bacteria exhibit a tree-like relationship with each other, often violated via horizontal gene transfer. As a result, the topology of each of their gene trees varies, so that their systematic taxonomy is constantly in flux; see for instance the paraphyletic Clostridium, Wiegel et al. (2006); Rainey et al. (2009). Their OTU proxies, typically the 16S rRNA gene, also exhibit a tree-like structure with patterns of cycles at deep taxon classification. Our resolution depends on the length of the 16S rRNA region, the degree of horizontal gene transfer, and the reliability of the reads.

In many problems, it is important to develop a means of selecting the OTUs having dominating roles in the microbial systems at hand. The relationship between OTUs is such that some of them may represent the same type of bacteria. Alternatively, some
spurious OTUs are artifacts of the laboratory sequencing protocol. Rather than select OTUs on their individual merits, we want to select them based on their group affiliations.

To facilitate comparison between microbiomes, sequences are grouped into OTUs according to some similarity criterion. This is a data-driven proxy to species delineation; we use OTUs in an operational manner. OTUs consist of sequences which are phylogenetically close. Phylogenetic proximity between OTUs is determined by the inter-OTU phylogenetic divergence of the constituent sequences. This provides a structural relationship between the OTUs. One manner of presenting this structure is to assign phylotypes to the OTUs, sequence-based consensus taxonomical classification of the clusters. This induces a rooted tree hierarchy between the OTUs. It is this structure we exploit in developing the phylogenetic LASSO, or $\Phi$-LASSO.

### 4.1 Generalized linear models and the $\Phi$-LASSO

As motivation, let us review the hierarchial HLASSO, presented in Zhou and Zhu (2010) in the context of penalized least-squares. Consider the linear model

$$ Y = X\beta + \epsilon, $$

(4.1.1)

for response $Y$, covariates $X$, parameter vector $\beta \in \mathbb{R}^p$, and error $\epsilon$. Suppose the covariates $X$ may be assigned to $k$ mutually exclusive groups $K_j$, $j = 1, \ldots, k$. Let $\beta_j \in \mathbb{R}^{p_j}$ be the subvector of $\beta$ whose coefficients correspond to the covariates in $K_j$, $|K_j| = p_j$, where $|\cdot|$ denotes set cardinality. We have $p = \sum_{j=1}^{k} p_j$.

We can decompose $\beta$ by $\beta_j = d_j \alpha_j$ where $d_j \geq 0$ and $\alpha_j \in \mathbb{R}^{p_j}$. Clearly this decomposition is not unique, but this does not ultimately matter, see Lemma 4.3.2 below.
Let \( d = (d_1, \ldots, d_k) \in \mathbb{R}^k \), \( \alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^p \), and define \( \varphi \) as the mapping given by 
\[(d, \alpha) \mapsto \beta.\]

We define the HLASSO estimator \( \hat{\beta} \) of (4.1.1) via \( \hat{\beta} = \varphi(\hat{d}, \hat{\alpha}) \) where \((\hat{d}, \hat{\alpha})\) maximizes the penalized least squares function
\[
-\frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i \cdot \varphi(d, \alpha))^2 - \lambda_1 \sum_{j=1}^{k} d_j - \lambda_2 \sum_{j=1}^{k} ||\alpha_j||_1,
\]
where the tuning parameters \( \lambda_1, \lambda_2 > 0 \) are fixed, \( \cdot \) denotes dot product, and \( || \cdot ||_q \) is the usual \( l_q \)-norm for \( q > 0 \). Here we penalize the groups \( K_j \) by the middle term \( \lambda_1 \sum_{j=1}^{k} d_j \) and the individual coefficients by the third term \( \lambda_2 \sum_{j=1}^{k} ||\alpha_j||_1 \).

It is shown in Zhou and Zhu (2010) that the penalty values \( \lambda_1, \lambda_2 \) redistribute geometrically so that we may replace them by a common coefficient \( \lambda = \sqrt{\lambda_1 \lambda_2} \). This leads to the result that their regularization is equivalent to maximizing
\[
-\frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i \cdot \beta)^2 - 2\lambda \sum_{j=1}^{k} \sqrt{||\beta_j||}.\]

The HLASSO is thus a nonconvex variant of the group LASSO, Yuan and Lin (2006).

While Zhou and Zhu (2010) directly treats least squares, their results generalize readily to arbitrary likelihood functions. They obtain the Oracle property by assuming local asymptotic normality, see Section 4.3.1. Below, we provide the generalization to accommodate increased depth to the hierarchy by framing it in terms of a taxonomy.

### 4.1.1 The \( \Phi \)-LASSO

Our goal now is to create a hierarchical penalization scheme where there are multiple competing ways of grouping covariates. If these groupings have a nesting property,
we can represent this as a tree, otherwise the graphical representation has cycles. In light of the above discussion, we use the convex log-likelihood function $\ell$ \textit{in lieu} of least squares in the sequel. We frame variable selection in terms of OTU selection in metagenomic analysis, but stress that the method easily accommodates other ‘-omic’ data forms, be they from metabolomics, proteomics, or transcriptomics. Further, as life betrays the ability to transmit genetic information horizontally, our method accommodates taxonomies as degenerate as the tree-of-life itself. In particular, this allows us to consider overlapping classifications of the variables, extending analysis to, say, active metabolic pathways of microbial systems or human tissues. We introduce some terminology and notation to help bridge the mathematics and systematics at hand.

\textbf{Definition 4.1.1} Let $X_1, \ldots, X_p$ be the $p$ column vectors of the design matrix $X \in \mathbb{R}^{n \times p}$. A \textbf{taxon} $\tau$ (plural \textbf{taxa}) is a subset of the indices $I = \{1, \ldots, p\}$ and $X_\tau = [X_j]_{j \in \tau}$ is the corresponding submatrix of $X$. A \textbf{taxon level} is a collection of pairwise disjoint taxa $\tau_k, k = 1, \ldots, K$ whose union is $I$.

As an example, consider Table 4.1, below. We find that taxa Bacilli = \{3, 4, 5, 6, 7, 8\} and \textit{Enterococcaceae} = \{3, 4, 5\}. The collection \{Actinobacteria, Bacilli, Clostridia\} is a taxon level. In general, we subdivide the indices into taxa at $T + 1$ taxon levels. We denote the $k$-th taxon of the $t$-th taxon level by $\tau^t$ or $\tau^t_k$.

\textbf{Definition 4.1.2} Suppose we have a collection $\mathcal{T}$ of $(T+1)$ taxon levels, where the $(T+1)$-th taxon level consists of singletons, $\{\{j\}\}_{j=1}^d$. We call $\mathcal{T}$ a \textbf{taxonomy}.

There will be times where we wish to refer to those indices that belong to specific taxa at each level. We have:
Table 4.1: An example taxonomy generated for thirteen OTUs using five taxon levels.

<table>
<thead>
<tr>
<th>Index</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>OTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Bifidobacteriales</td>
<td>Bifidobacteriaceae</td>
<td>OTU₁</td>
</tr>
<tr>
<td>2</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Bifidobacteriales</td>
<td>Bifidobacteriaceae</td>
<td>OTU₂</td>
</tr>
<tr>
<td>3</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Enterococcaceae</td>
<td>OTU₃</td>
</tr>
<tr>
<td>4</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Enterococcaceae</td>
<td>OTU₄</td>
</tr>
<tr>
<td>5</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Enterococcaceae</td>
<td>OTU₅</td>
</tr>
<tr>
<td>6</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Lactobacillaceae</td>
<td>OTU₆</td>
</tr>
<tr>
<td>7</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Lactobacillaceae</td>
<td>OTU₇</td>
</tr>
<tr>
<td>8</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillales</td>
<td>Lactobacillaceae</td>
<td>OTU₈</td>
</tr>
<tr>
<td>9</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Clostridiales</td>
<td>OTU₉</td>
</tr>
<tr>
<td>10</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Clostridiales</td>
<td>OTU₁₀</td>
</tr>
<tr>
<td>11</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>OTU₁₁</td>
</tr>
<tr>
<td>12</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>OTU₁₂</td>
</tr>
<tr>
<td>13</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>OTU₁₃</td>
</tr>
</tbody>
</table>

**Definition 4.1.3** Let \( L = (\tau^1, ..., \tau^T) \) be a \( T \)-tuple where taxon \( \tau^t \) belongs to taxon level \( t \). We refer to \( L \) as a **lineage**, with associated indices \( J = \cap_{t=1}^{T} \tau^t \).

We write \( X_L \) in lieu of \( X_J \) and define \(|L| := |J|\). When we wish to make it clear we are referring to a lineage’s taxon at a particular taxon level, we write \( L^t \), so that \( L \) may be re-written \( L = (L^1, ..., L^T) \).

In our example, the lineages may be read directly off Table 4.1, of which there are five. Figure 4.1 presents the taxonomy more intuitively. Each branch of the tree represents a taxon while each horizontal row represents a taxon level, numbered 1 through 5. Each lineage is a directed path from the root to a branch. The definitions are more general, however, in that there is no nesting property assumed or implied, i.e. it is not necessary that the \((k+1)\)th taxon level be a subdivision of the \(k\)th.

Consider a generalized linear model with a known link function \( g \), so that \( \mathbb{E}(Y) = \)
Figure 4.1: A graphical representation of the taxonomy as well as an illustration of the decomposition of the parameters $\beta$ following the taxonomy in Table 4.1.

$g^{-1}(X\beta)$. We decompose $\beta$ by $\beta_L = d_L\alpha_L$, where $\alpha_L \in \mathbb{R}^{\left|L\right|}$ as before, but now $d_L = \prod_{t=1}^{T}d_{Lt}$, $d_L^t \geq 0$ for $t = 1, ..., T$. We write this decomposition as $(D, \alpha)$. Let $\varphi$ be the map $(D, \alpha) \mapsto \beta$. This decomposition is illustrated in parentheses in Figure 4.1: the coefficients $d_L$ are recovered by multiplying the terms in a lineage. We extend the optimization criterion to:

$$\ell(\varphi(D, \alpha); Y, X) - \sum_{t=1}^{T} \lambda_t \sum_{k=1}^{K^t} d_k^t - \lambda_{T+1}||\alpha||_1$$  \hspace{1cm} (4.1.2)$$

where as usual $\ell$ is the log-likelihood, and $\lambda_t > 0$ for $t = 1, ..., T + 1$. The centre double sum is the groups penalty and addresses the increased depth of our taxonomy in contrast to Zhou and Zhu (2010) and $K^t$ is the number of taxa in taxon level $t$. When the lineage consists of a single taxon level $T = 1$, this reduces to the criterion in Zhou and Zhu (2010).

There is an affluence of tuning parameters. As an extension of Zhou and Zhu (2010), Lemma 4.3.1 shows that criterion (4.1.2) is equivalent to the single tuning parameter
criterion

\[ \ell(\varphi(D, \alpha); Y, X) - \sum_{t=1}^{T} \sum_{k=1}^{K_t} d_{t}^k - \lambda \|\alpha\|_1 \quad . \tag{4.1.3} \]

Let \( d^t \) denote the vector of groups coefficients of the \( t \)-th taxon level. Intuitively, as the tuning parameters \( \lambda_t \) are redistributed, \( d^t, t = 1, \ldots, T \) and \( \alpha \) change in geometric response due to the relationship \( \varphi(D, \alpha) = \beta \). Choosing \( \lambda = \lambda_t, t = 1, \ldots, T + 1 \), we need only concern ourselves with one tuning parameter.

4.1.2 \( \Phi \)-LASSO algorithm

The algorithm used to obtain the \( \Phi \)-LASSO estimate relies on iterative adaptive reweighting, deriving from the reformulation of the \( \Phi \)-LASSO in (4.1.3). Let \( \psi: \beta \mapsto (D, \alpha) \) be the map from \( \beta \) to the unique maximizer of (4.1.3) over \( \varphi^{-1}(\beta) \), guaranteed by Lemma 4.3.2. In this way \( D \) and \( \alpha \) may be viewed as projections onto the first and second elements of \( \varphi^{-1}(\beta) \). Let \( \Lambda \) be a positive collection of tuning parameters \( \lambda \). The pseudo-code is spelled out below. We remark that convergence is typically achieved quickly, so that the bottlenecks are the weighted LASSO problem and calculation of \( w \). The code is provided in Appendix A. The justification for this algorithm is heuristic, as it has been shown that initializing gradient descent or LLA from a LASSO solution attains local solutions satisfying the oracle property with asymptotic probability 1, Zhang (2010b); Zhang and Zhang (2012); Fan et al. (2014).
Algorithm 1 \( \Phi \)-LASSO

procedure PHYLASSO

for each \( \lambda \in \Lambda \) do

obtain initial LASSO estimate \( \hat{\beta}^{0(\lambda)} \)

for \( k \in \mathbb{N} \) do

\( w = \varphi(D(\hat{\beta}^{(k-1)(\lambda)}), 1) \)

solve weighted LASSO for \( \hat{\beta}^{k(\lambda)} \) with weights \( w^{-1} \)

break if \( \Delta \hat{\beta}^{\lambda} < \) threshold

end for

end for

end procedure

4.2 Simulations

In this section, we report on simulations that approximate the covariance structure we may see in practice with respect to phylogenetic proximity of the OTUs. We address different questions with these. In simulation 1, we are interested in the performance of the \( \Phi \)-LASSO when the hyperparameter \( \lambda \) is already optimized. In simulation 2, we address the situation we encounter in practice, where we must cross-validate the tuning parameter.

4.2.1 Simulation 1

We consider \( T = 5 \) taxon levels where each level grows according to \( 4^k \), \( k = 0, 1, 2, 3, 4, 5, 6 \) and is balanced in the following sense: we have a single ‘phylum’ (\( k = 0 \)) followed by four ‘classes’ (\( k = 1 \)), each with four ‘orders’ (\( k = 2 \)), each with four ‘families’
(k = 3), each with four ‘genera’ (k = 4), each with four ‘species’ (k = 5). This results in 4096 OTUs (k = 6).

To each taxon we introduce taxon-wise covariation. This covariance structure is presented in Figure 4.2 for a single ‘order’ (256 OTUs) using a 10,000 point sample. The true parameters in the simulation are represented by two classes, with one class dominating the other 3:1.

Figure 4.2: Heatmap of covariance matrix for 10,000 point validation set from the tuning parameter step. Displayed is the submatrix corresponding to a single ‘order’.

We consider random Gaussian data generated from a 4096 covariate sparse linear model where 32 covariates have corresponding parameter $\beta_j = 2$ and the rest being zero. We present the corresponding pruned tree in Figure 4.3.

The species information is deliberately lost to simulate uncertainty in species assignment to OTUs, which mimics the current technology limitation in 16s rRNA sequencing, see Chakravorty et al. (2007). Thus the deepest taxon level used in the fit is the genus level. We introduce class, order, family, genus, and species-wise covariation $W_c, W_o, W_f, W_g$, and
Figure 4.3: Pruned taxonomy for simulations 1 and 2, from an initial 4096 and 16384 leaves, respectively.

$W_s$, sampled from normal distributions $N(0, 0.5^2)$, $N(0, 1^2)$, $N(0, 2^2)$, $N(0, 3^2)$, $N(0, 4^2)$, respectively, where $N(0, \sigma^2)$ denotes a normal distribution with mean 0 and variance $\sigma^2$. The individual predictors used are then defined as $X_j = (Z_j + W_p + W_c + W_f + W_g + W_s)/\sqrt{55.25}$ where $Z_j \sim N(0, 5^2)$ and $j$ belongs to lineage ($c, o, f, g, s$).

**Tuning parameter selection**

To select the tuning parameter for each $(n, \sigma)$ pair, we replicate 100 data sets $Y_i \sim N(X_i \beta, \sigma)$, $1 \leq i \leq n$. We select the tuning parameter $\lambda^{(n, \sigma)}$ minimizing across all models the MSPE (mean squared prediction error) for an independent 10,000 data point validation set generated from the same distribution. We also fit SCAD models using the same datasets, using the same validation set to select the appropriate tuning parameter.

**Performance**

To evaluate performance of our models, we consider four measures for 100 replicates: SSE (sum of squared error), MSPE, ‘recall’, and ‘precision’. Recall is defined as the proportion of covariates correctly selected relative to true parameters, $tp/(tp+fn)$, and precision is defined as the proportion of covariates correctly selected relative to total
covariates selected, \( \frac{tp}{tp+fp} \), where \( tp \) is true positive, \( fn \) is false negative and \( fp \) is false positive.

We compare the performance of the \( \Phi \)-LASSO to SCAD. We also consider OLS (ordinary least squares) for the oracle model where the exact \( \beta_j = 0 \) is known. MSPE for performance is evaluated using a new 10,000 data point validation set generated independently of the tuning validation set.

**Results: Tuning**

Figure 4.4 presents recall and precision against the tuning parameters. Also displayed is the MSPE curve, scaled to the unit interval \([0, 1]\). Note that the selected parameters \( \lambda \) lie in the middle of a relatively flat MSPE region. Perturbations of \( \lambda \) would not overly affect predictive performance. The same holds with respect to SSE (not shown). The shapes of the MSPE curves agree in large measure, aside from the null models, which have lower SSE than the sparsest estimators. Recall and precision are at odds, where for low sample size one must be sacrificed against the other. The selection of \( \lambda \) favours higher recall over precision; the coefficients for false positives may be very small and hence affect prediction in a minor way, but false negatives are a complete loss of a structural signal. There is an interesting dip in recall for sample sizes \( n \geq 100 \) in the region preceding the low, stable MSPE region where the estimators overfit the data. The clear separation of these two regions reflects well on the \( \Phi \)-LASSO stability.
Results: Performance

The results for SSE and MSPE are presented in Table 4.2 and those for recall and precision in Table 4.3. SCAD struggles with the covariance structure, performing poorly in estimation. It appears to perform well in MSPE. While SCAD struggles to recall the correct OTUs due to high correlation within species, to its credit it is able to select some related taxa, where the covariance structure leads to similar predictive performance but more flexibility in fitting the model. The $\Phi$-LASSO exhibits none of the SCAD’s difficulties. It quickly converges to agreement with the oracle estimator in estimation and prediction.

The $\Phi$-LASSO quickly approaches near perfect recall as sample size increases whereas SCAD becomes stalled at 25%. The $\Phi$-LASSO consistently improves in precision, dropping false positives. For SCAD, after an initial improvement in precision, it drops as it selects incorrect but related OTUs. When we present the $\Phi$-LASSO and SCAD estimates with a validation set consisting of uncorrelated covariates, there is negligible change in $\Phi$-LASSO performance, while the SCAD’s predictive performance matches its poor estimation performance.

4.2.2 Simulation 2

We consider $T = 6$ taxon levels where each level grows according to $4^k$, $k = 1, 2, 3, 4, 5, 6, 7$ and is balanced in the following sense: we have four ‘phyla’ ($k = 1$) followed by ‘class’ ($k = 2$), ‘order’ ($k = 3$), ‘family’ ($k = 4$), ‘genus’ ($k = 5$), and ‘species’ ($k = 6$) with four representative OTUs, for a total of 16,384 OTUs ($k = 7$).

To each taxon we introduce taxon-wise covariation similar to Simulation 1. The
true parameters in the simulation are represented by two classes within a single phylum, with one class dominating the other 3:1.

We consider random Gaussian data generated from a 16,384 covariate sparse linear model where 32 covariates have corresponding parameter sampled independently and uniformly from $U(0.5, 3.5)$ and the rest being zero. These parameters are fixed over all replicates. We present the corresponding pruned tree in Figure 4.3.

The species information is again deliberately lost to simulate uncertainty in species assignment to OTUs. Thus the deepest taxon level used in the fit is the genus level. We introduce a phylum, class, order, family, genus, and species-wise covariation $W_p, W_c, W_o, W_f, W_g, \text{ and } W_s$, sampled from normal distributions $N(0, 0.5^2), N(0, 1^2), N(0, 2^2), N(0, 3^2), N(0, 4^2), N(0, 5^2)$, respectively. The predictors used are then defined as $X_j = (Z_j + W_p + W_c + W_f + W_g + W_s) / \sqrt{55.25}$ where $Z_j \sim N(0, 6^2)$ and $j$ belongs to lineage $(p, c, o, f, g, s)$.

**Tuning parameter selection**

To select the tuning parameters, we replicate 100 data sets $Y_i \sim N(X_i \beta, \sigma)$, $1 \leq i \leq n$ and perform 5-fold cross-validation optimizing MSPE. We select the tuning parameters $\lambda^{(n, \sigma, r)}$ minimizing the k-fold cross-validation error for each sample size $n$, noise level $\sigma$, and replicate $r$. We also fit MCP and SCAD models using the same datasets, using the same validation method to select the appropriate tuning parameter.
Performance

To evaluate performance of our models, we consider four measures for 100 replicates: SSE, MSPE, ‘recall’, and ‘precision’. We compare the performance of the $\Phi$-LASSO to MCP and SCAD. We also consider OLS for the oracle model where the exact $\beta_j = 0$ are known. MSPE for performance is evaluated using a 10,000 data point validation set generated independently tuning and fitting data.

Results

The results for SSE and MSPE are presented in Table 4.4, and those for recall and precision in Table 4.5. The results of simulation 2 are qualitatively identical to those for simulation 1. The $\Phi$-LASSO estimator quickly converges to oracle performance, while MCP and SCAD lag considerably behind. MCP and SCAD achieve perfect precision by underfitting the data, stalling at around 25% recall. The $\Phi$-LASSO outperforms in terms of recall and consistently improves in recall and precision.

4.2.3 Additional simulations

We comment on some additional simulations. In the low dimensional, $p < n$, regime with simple sample-wise covariation structure, SCAD performs moderately better than the $\Phi$-LASSO, with the adaptive LASSO trailing behind. This is consistent across sample size and noise level. For alternative $p > n$ scenarios ($p = 1000$), the $\Phi$-LASSO performs well, although experiences some difficulty with precision for $n = 250$. This is qualitatively different than the issue with SCAD in the previous section, as recall is nearly
perfect. The false positives generally correspond to relatively small coefficients, so that it appears to result from early exit from the fitting algorithm.

Figure 4.4: Simulation 1. Median recall (solid) and precision (dashed) for Φ-LASSO against log(λ) for sample sizes $n = 50, 100, 150$. Included is the median MSPE curve scaled by largest median (dotted). The ‘+’ indicates the chosen (log) tuning parameter.
### Table 4.2: Simulation 1. Estimation and prediction error for the oracle estimator (OLS), \( \Phi \)-LASSO, and SCAD. Presented are the mean error (standard error).

<table>
<thead>
<tr>
<th>( n )</th>
<th>OLS</th>
<th>( \Phi )-LASSO</th>
<th>SCAD</th>
<th>OLS</th>
<th>( \Phi )-LASSO</th>
<th>SCAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>88.10(46.2)</td>
<td>167.00(62.50)</td>
<td>621.00(230.00)</td>
<td>3.20(1.21)</td>
<td>14.80(4.48)</td>
<td>9.24(0.56)</td>
</tr>
<tr>
<td>100</td>
<td>19.80(6.05)</td>
<td>37.70(12.00)</td>
<td>384.00(5.28)</td>
<td>1.47(0.14)</td>
<td>2.43(0.18)</td>
<td>8.66(0.38)</td>
</tr>
<tr>
<td>150</td>
<td>11.10(3.69)</td>
<td>15.90(6.14)</td>
<td>383.00(4.47)</td>
<td>1.25(0.07)</td>
<td>1.49(0.18)</td>
<td>8.40(0.24)</td>
</tr>
<tr>
<td>200</td>
<td>8.05(2.48)</td>
<td>10.70(4.29)</td>
<td>379.00(4.35)</td>
<td>1.18(0.05)</td>
<td>1.29(0.10)</td>
<td>8.27(0.20)</td>
</tr>
<tr>
<td>250</td>
<td>6.25(1.77)</td>
<td>7.54(2.45)</td>
<td>380.00(3.36)</td>
<td>1.13(0.03)</td>
<td>1.19(0.06)</td>
<td>8.27(0.20)</td>
</tr>
</tbody>
</table>

### Table 4.3: Simulation 1. Recall and precision for the \( \Phi \)-LASSO and SCAD. Presented are the mean (standard error).

<table>
<thead>
<tr>
<th>( n )</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \Phi )-LASSO</td>
<td>SCAD</td>
</tr>
<tr>
<td>50</td>
<td>0.64(0.11)</td>
<td>0.18(0.07)</td>
</tr>
<tr>
<td>100</td>
<td>0.92(0.04)</td>
<td>0.25(0.00)</td>
</tr>
<tr>
<td>150</td>
<td>0.98(0.02)</td>
<td>0.25(0.00)</td>
</tr>
<tr>
<td>200</td>
<td>0.99(0.01)</td>
<td>0.25(0.00)</td>
</tr>
<tr>
<td>250</td>
<td>0.99(0.01)</td>
<td>0.25(0.00)</td>
</tr>
</tbody>
</table>

### Table 4.4: Simulation 2. Estimation and prediction error for the oracle estimator (OLS), \( \Phi \)-LASSO, and SCAD. Presented are the mean error (standard error).

<table>
<thead>
<tr>
<th>( n )</th>
<th>SSE</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>( \Phi )-LASSO</td>
</tr>
<tr>
<td>50</td>
<td>132(61.4)</td>
<td>320(145)</td>
</tr>
<tr>
<td>100</td>
<td>35.2(12.2)</td>
<td>92.6(33.6)</td>
</tr>
<tr>
<td>150</td>
<td>18.6(5.76)</td>
<td>48.0(48.9)</td>
</tr>
<tr>
<td>200</td>
<td>13.2(4.17)</td>
<td>31.6(10.4)</td>
</tr>
<tr>
<td>250</td>
<td>10.3(3.39)</td>
<td>26.1(8.05)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( n )</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
</tr>
<tr>
<td>50</td>
<td>3.02(0.92)</td>
</tr>
<tr>
<td>100</td>
<td>1.52(0.17)</td>
</tr>
<tr>
<td>150</td>
<td>1.28(0.08)</td>
</tr>
<tr>
<td>200</td>
<td>1.20(0.06)</td>
</tr>
<tr>
<td>250</td>
<td>1.16(0.04)</td>
</tr>
</tbody>
</table>

Table 4.2: Simulation 1. Estimation and prediction error for the oracle estimator (OLS), \( \Phi \)-LASSO, and SCAD. Presented are the mean error (standard error).

Table 4.3: Simulation 1. Recall and precision for the \( \Phi \)-LASSO and SCAD. Presented are the mean (standard error).

Table 4.4: Simulation 2. Estimation and prediction error for the oracle estimator (OLS), \( \Phi \)-LASSO, and SCAD. Presented are the mean error (standard error).
Table 4.5: Simulation 2. Recall and precision for the $\Phi$-LASSO and SCAD. Presented are the mean (standard error).

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Phi$-LASSO Recall</th>
<th>SCAD Recall</th>
<th>MCP Recall</th>
<th>$\Phi$-LASSO Precision</th>
<th>SCAD Precision</th>
<th>MCP Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.40(0.13)</td>
<td>0.15(0.05)</td>
<td>0.13(0.04)</td>
<td>0.37(0.14)</td>
<td>0.41(0.22)</td>
<td>0.74(0.24)</td>
</tr>
<tr>
<td>100</td>
<td>0.75(0.08)</td>
<td>0.24(0.03)</td>
<td>0.23(0.02)</td>
<td>0.68(0.09)</td>
<td>0.90(0.17)</td>
<td>1.00(0.01)</td>
</tr>
<tr>
<td>150</td>
<td>0.84(0.07)</td>
<td>0.25(0.02)</td>
<td>0.24(0.02)</td>
<td>0.79(0.09)</td>
<td>0.99(0.03)</td>
<td>1.00(0.00)</td>
</tr>
<tr>
<td>200</td>
<td>0.90(0.05)</td>
<td>0.26(0.03)</td>
<td>0.24(0.01)</td>
<td>0.86(0.07)</td>
<td>1.00(0.01)</td>
<td>1.00(0.00)</td>
</tr>
<tr>
<td>250</td>
<td>0.92(0.05)</td>
<td>0.26(0.03)</td>
<td>0.24(0.02)</td>
<td>0.90(0.06)</td>
<td>1.00(0.00)</td>
<td>1.00(0.00)</td>
</tr>
</tbody>
</table>

4.3 Theoretical Results

In this section, we consider the general objective function,

$$Q^\ast(\lambda_1, \ldots, \lambda_{T+1}, D, \alpha) = \ell(\varphi(D, \alpha)) - \sum_{t=1}^{T} \lambda_t ||D_t||_q^q - \lambda_{T+1}||\alpha||_q^q$$

where $\ell$ is the known but arbitrary log-likelihood, the remainder is the penalization, and $|| \cdot ||_q$ is the $l_q$-norm, $q > 0$. We will also make use of the notation ‘$\cdot$’ for matrix transpose, ‘$\ll$', ‘$\ll_p$’ to mean ‘big oh’ and ‘big oh in probability’, respectively, $o$, $o_p$ to mean ‘little oh’ and ‘little oh in probability’, respectively, ‘$\asymp$’ to mean the ratio of two sequences converge to a positive constant, ‘$\rightsquigarrow$’ to mean convergence in distribution, and ‘$\wedge$’ to mean minimum.

All proofs are provided in Section 4.4.

The first result allows us to consider a single penalty parameter, that is, set $\lambda_t = \lambda$ for all $t = 1, \ldots, T + 1$. Let $Q_1 = Q^\ast(\lambda_1, \ldots, \lambda_{T+1}, \cdot, \cdot, \cdot)$, $Q_2 = Q^\ast(1, \ldots, 1, \lambda_{T+1} \prod_{t=1}^{T} \lambda_t, \cdot, \cdot)$.

Lemma 4.3.1 (Equivalence of optimization) Let $\lambda_t > 0$, $t = 1, \ldots, T + 1$ be fixed. Then $(D^1, \alpha^1)$ is a local maximizer of $Q_1$ if and only if $(D^2 = (\lambda_t^{1/q}d_t^1), \alpha^2 = \alpha^1/\prod_{t=1}^{T} \lambda_t^{1/q})$ is a local maximizer of $Q_2$ and hence $\varphi(D^1, \alpha^1) = \varphi(D^2, \alpha^2)$.

The proof is similar to that of Lemma 1 in Zhou and Zhu (2010), where least-squares with
$l_1$-regularization is considered.

In the sequel we assume that $\lambda_t = \lambda$ for all $t = 1, \ldots, T + 1$. This next result provides a relationship between the individual effects parameters and the group coefficients. Intuitively, the mass distributes itself geometrically to minimize the penalty terms. Through abuse of notation, we allow $\varphi^{-1}(\beta)$ to mean the inverse image of $\{\beta\}$ with respect to $\varphi$.

**Lemma 4.3.2 (Mass Equilibrium)** Let $(D, \alpha) \in \varphi^{-1}(\beta)$. Then $(D, \alpha)$ is a global maximizer of $Q^*$ over $\varphi^{-1}(\beta)$ if and only if $d_{\tau_k}^L = \sum_{L \neq \tau_k} \|\alpha_L\|_q^q = \|\alpha_{\tau_k}\|_q^q$.

This conservation of mass result immediately leads to the following.

**Corollary 4.3.3** The relationship $d_{\tau_k}^L = \sum_{L \neq \tau_k} \|\alpha_L\|_q^q$ from Lemma 4.3.2 identifies the unique maximizer of $Q^*$ over $\varphi^{-1}(\beta)$.

Thus $\beta_j$ is an order $T + 1$ polynomial in terms of $\alpha_1, \ldots, \alpha_p$.

**Definition 4.3.1** Let $\psi$ be the map $\beta \mapsto (D, \alpha)$ where $(D, \alpha)$ is the unique optimizer of $Q^*|_{\varphi^{-1}(\beta)}$. We refer to $\psi$ as the partial inverse.

**Corollary 4.3.4** Suppose the taxonomy has two levels, $T = 1$. Then the partial inverse $\psi(\beta) = (D, \alpha)$ is characterized by: if $\beta_L = 0$ then $d_L = 0$ and $\alpha_L = 0$, if $\beta_L \neq 0$ then $d_L = \sqrt{\|\beta_L\|_q}$ and $\alpha_L = \beta_L / \sqrt{\|\beta_L\|_q}$.

Corollary 4.3.4 is a generalization of Theorem 1 in Zhou and Zhu (2010).
4.3.1 LAN conditions and the Oracle

To ensure local asymptotic normality of the MLE for increasing number of parameters, we adopt the regularity conditions from Lehman and Casella (2003) as in Zhou and Zhu (2010):

(A1) For all \( n \), the observations \((X_i, Y_i), i = 1, \ldots, n\), are independently and identically distributed according to the density \( f_n(X_i, Y_i; \beta_n) \), where \( f_n \) has common support and the model is identifiable. Further,

\[
\mathbb{E}_{\beta_n} \left[ \frac{\partial \log f_n}{\partial \beta_{nL_j}} \right] = 0 \quad \forall L, j = 1, \ldots, |L|
\]

\[
\mathcal{I}_{L, K_k}(\beta_n) = \mathbb{E}_{\beta_n} \left[ \frac{\partial}{\partial \beta_{nL_j}} \log f_n \cdot \frac{\partial}{\partial \beta_{nK_k}} \log f_n \right] = -\mathbb{E}_{\beta_n} \left[ \frac{\partial^2}{\partial \beta_{nL_j} \partial \beta_{nK_k}} \log f_n \right].
\]

(A2) The Fisher information matrix \( \mathcal{I}(\beta_n) = (\mathcal{I}_{L, K_k}(\beta_n)) \) is positive definite with bounds

\[
0 < C_1 < \min \sigma(\mathcal{I}(\beta)) \leq \max \sigma(\mathcal{I}(\beta)) < C_2 < \infty
\]

where \( \sigma(X) \) is the point spectrum of a matrix \( X \),

\[
\mathbb{E}_{\beta_n} \left[ \frac{\partial}{\partial \beta_{nL_j}} \log f_n \cdot \frac{\partial}{\partial \beta_{nK_k}} \log f_n \right]^2 < C_3 < \infty
\]

\[
\mathbb{E}_{\beta_n} \left[ \frac{\partial^2}{\partial \beta_{nL_j} \partial \beta_{nK_k}} \log f_n \right]^2 < C_4 < \infty.
\]

(A3) There exists an open subset \( \omega_n \subset \Omega_n \subset \mathbb{R}^{p_n} \), \( p_n \) the number of parameters, containing the true parameter point \( \beta_n \) such that for almost all \((X_i, Y_i)\), \( \frac{\partial^3 f_n}{\partial \beta_{L_j} \partial \beta_{K_k} \partial \beta_{L_i}} \) is defined for all \( \beta \in \omega_n \). Further, there exist functions \( M_{nL_j K_k J_l} \) such that

\[
\left| \frac{\partial^3 f_n}{\partial \beta_{L_j} \partial \beta_{K_k} \partial \beta_{L_i}} \right| \leq M_{nL_j K_k J_l}(X_i, Y_i)
\]
for all $\beta \in \omega_n$ and
\[ \mathbb{E}_{\beta_n} \left[ M^2_{nLjK_iL}(X_i, Y_i) \right] < C_5 < \infty. \]

We refer to the conditions (A1)-(A3) as the local asymptotic normality (LAN) conditions and note that they are not excessively restrictive and fall within the usual framework of this type of analysis.

Consider the parameters $\beta_n \in \mathbb{R}^{p_n}$ where $p_n$ grows with $n$, and their partial inverse $\psi(\beta_n) = (D_n, \alpha_n)$. Let $w_j(\beta_n) = 1$ if $\varphi_j(D_n(\beta_n), 1) = 0$, $\varphi_j(D_n(\beta_n), 1)$ otherwise. Then assuming the relationship from Lemma 4.3.2 for $q = 1$, we have the following equivalent expression,
\[
Q_n(\lambda_n, \beta_n) = Q^*(n\lambda_n/(T + 1), ..., n\lambda_n/(T + 1), D_n, \alpha_n)
\]
\[
= \ell_n(\beta_n) - n\lambda_n \sum_{L} \frac{||\beta_{n,L}||}{w_{L_1}}. \tag{4.3.1}
\]

For convenience we write $D_n(\beta_n)$ for the projection of $\psi(\beta_n)$ onto $D_n$.

We have the following consistency results for the $\Phi$-LASSO.

**Theorem 4.3.5 (Consistency)** Assume that the distribution satisfies the LAN conditions. If $p_n^4 = o(n)$ and $\lambda_n \ll n^{-1/2}$, then there exists a $\gamma_n$-consistent local maximizer $\hat{\beta}_n$ of $Q_n$, where $\gamma_n = \sqrt{p_n}(n^{-1/2} + \lambda_n)$.

A simple choice is $\lambda_n \asymp n^{-1/2}$. We find this provides us with taxon selection consistency.

**Theorem 4.3.6 (Taxon selection consistency)** Assume that the distribution satisfies the LAN conditions. If $p_n^{(T+2)/4} = o(n)$ and $\lambda_n \asymp n^{-1/2}$, then there exists a $\sqrt{n/p_n}$-consistent local maximizer $\hat{\beta}_n$ of (4.3.1) so that $\lim_{n \to \infty} \mathbb{P}\{\delta(\hat{\beta}_{nL}) = \delta(\beta^0_{nL})\} = 1$, the group sparsity property.
For the case of two taxon levels, $T = 1$, the above result is Theorem 2, Zhou and Zhu (2010). Using an adaptive modification to their loss function analogous to Zou (2006), Zhou and Zhu (2010) obtain the full oracle property. We obtain the oracle property in an alternative manner, by a simple modification of the taxonomy of the variables. The full proofs are quite involved and are relegated to Section 4.4. Since we have now shown that consistent taxon selection holds for arbitrary number of taxon levels, we have the full oracle property in the following result using a simple modification of our taxonomy. This is possible as we are able to incorporate $T$, arbitrary, taxon levels.

**Theorem 4.3.7 (Oracle Property)** In addition to the LAN conditions, suppose that the taxonomy contains an additional taxon level identical to the singleton level and $p_n^5 = o(n)$. Then $\hat{\beta}_n$ is a $\sqrt{n/p_n}$-consistent local maximizer of (4.3.1) satisfying

(i) model consistency $\mathbb{P}\{\mathcal{A}_n = \mathcal{A}_n^0\} \rightarrow 1$ as $n \rightarrow \infty$,

(ii) asymptotic normality

$$\sqrt{nA_nT_n^{1/2}}(\beta_n^0 - \beta_{n,\mathcal{A}_n}^0) \sim N(0, \Sigma),$$

where $A_n$ is an $r \times \mathcal{A}_n$ matrix such that $A_nA_n' \rightarrow \Sigma$, a positive semidefinite matrix, and $\mathcal{I}_n(\beta_n^0)$ is the Fisher information matrix evaluated at $\beta_n^0$, as $n \rightarrow \infty$.

**Remark 4.3.8** Alternatively, we can obtain the oracle property from the following modification of the objective function,

$$\ell(\varphi(D, \alpha); Y, X) - \sum_{t=1}^{T} \sum_{k=1}^{K_t} d_k^t - \lambda ||\alpha||_{\frac{1}{2}}^\gamma,$$

where we have replaced the LASSO penalty on $\alpha$ by a Bridge penalty with $\gamma = 1/2$. The $l_{1/2}$-penalty yields an estimator equivalent to that obtain in Theorem 4.3.7.
A simple consequence of Theorem 4.3.7 is that all estimators obtained with an $l_{1/q}$ penalty, $q \in \mathbb{N}, q > 1$, have the oracle property.

**Corollary 4.3.9** Consider the estimator $\hat{\beta}_n = \arg\max_\beta \{ \ell_n(\beta) - \lambda_n||\beta||_{1/T+1}^{1/T+1} \}$. Then $\hat{\beta}_n$ has the oracle property.

This result follows immediately if we consider a taxonomy in which all taxa are singletons.

### 4.4 Proofs

In this section we provide all proofs.

**Proof of Lemma 4.3.1.** Clearly $\varphi(D^1, \alpha^1) = \varphi(D^2, \alpha^2)$.

Then

$$Q_1(D, \alpha) = \ell(\varphi(D, \alpha)) - \sum_{t=1}^{T} \lambda_t ||D_t||_q - \lambda_{T+1}||\alpha||_q$$

$$= \ell(\varphi(D, \alpha)) - \sum_{t=1}^{T} ||\lambda_t^{1/q}D_t||_q - \lambda_{T+1} \left( \prod_{t=1}^{T} \lambda_t \right) \left\| \alpha / \prod_{t=1}^{T} \lambda_t^{1/q} \right\|_q$$

$$= Q_2((\lambda_t^{1/q}d_t), \alpha / \prod_{t=1}^{T} \lambda_t^{1/q}) \quad (4.4.1)$$

(\implies) Let $(D^1, \alpha^1)$ be a local maximizer of $Q_1$. Thus there exists a $\delta > 0$ such that if $(D', \alpha')$ satisfies $\sum_t ||d'_t - d_t|| + ||\alpha' - \alpha^1|| < \delta$, then $Q_1(D', \alpha') \leq Q_1(D^1, \alpha^1)$. By the identity (4.4.1),

$$Q_2(D'' = (\lambda_t^{1/q}d'_t), \alpha'' = \alpha' / \prod_{t=1}^{T} \lambda_t^{1/q}) = Q_1(D', \alpha') \leq Q_1(D^1, \alpha^1) = Q_2(D^2, \alpha^2).$$

As this holds for $(D'', \alpha'')$ in a neighbourhood of $(D^2, \alpha^2), (D^2, \alpha^2)$ is a local maximizer of $Q_2$. 

The converse is similar. □

From hereon, we assume \(\lambda_t = \lambda, t = 1, \ldots, T + 1\). Before proving Lemma 4.3.2, we recall the Karush-Kuhn-Tucker (KKT) conditions from convex analysis, here taken from Rockafellar (1970).

**Definition 4.4.1** Let \(C \neq \emptyset\) be a convex subset of \(\mathbb{R}^n\). Let \(f_i : C \rightarrow \mathbb{R}\) be convex functions on \(C\) for \(0 \leq i \leq r\) and affine functions on \(C\) for \(r + 1 \leq i \leq m\). Consider the following problem,

\[
\begin{align*}
\text{(P)} & \quad \text{minimize } f_0(x) \\
& \quad \text{subject to } f_i(x) \leq 0 \quad (1 \leq i \leq r) \\
& \quad \quad f_i(x) = 0 \quad (r + 1 \leq i \leq m)
\end{align*}
\]

We call \((P)\) an ordinary convex program.

**Lemma 4.4.1 (KKT Conditions, Rockafellar (1970))** Let \((P)\) be an ordinary convex program. Let \(\mu \in \mathbb{R}^m, x \in \mathbb{R}^n\). In order for \(\mu\) to be a KKT vector for \((P)\) and \(x\) an optimal solution to \((P)\), it is necessary and sufficient that \((\mu, x)\) be a saddle-point of the Lagrangian of \((P)\). Moreover, this condition holds if and only if \(x\) and the components \(\mu_i\) of \(\mu\) satisfy

(i) \(\mu_i \geq 0, f_i(x) \leq 0, \text{ and } \mu_i f_i(x) = 0 (1 \leq i \leq r)\);

(ii) \(f_i(x) = 0 (r + 1 \leq i \leq m)\);

(iii) \(0 \in [\partial f_0(x) + \sum_{i=1}^{m} \mu_i \partial f_i(x)], \text{ the subgradient of the Lagrangian at } x\).

**Proof of Lemma 4.3.2.** Note that the log-likelihood component of the loss function depends on \((D, \alpha)\) through \(\varphi(D, \alpha) = \beta\), so that it remains constant over \(\varphi^{-1}(\beta)\). Without
loss of generality, assume $\alpha > 0$, since if a component is negative, we can multiply the variable in question by $-1$, and if it is zero we can exclude it from the current analysis, as it remains fixed. We further assume $d_\tau > 0$ since if it is zero, then again, it is fixed.

We have the following optimization program,

$$
\begin{align*}
\text{(P_1)} \quad & \text{minimize} & & \sum_i d^q_i + \|\alpha\|^q \\
\text{subject to} & & & -d_\tau < 0 \\
& & & (\Pi d^L_\tau) \alpha_{L_j} = \beta_{L_j} \quad (\forall \tau, L_j).
\end{align*}
$$

Let $x = (x_\tau) = (\ln d_\tau)$, $y = (y_{L_j}) = (\ln \alpha_{L_j})$, and $B_{L_j} = \ln \beta_{L_j}$. Then (P_1) is equivalent to the following ordinary convex program,

$$
\begin{align*}
\text{(P_2)} \quad & \text{minimize} & & f(x, y) = \sum \tau e^{q x_\tau} + \sum L_j e^{q y_{L_j}} \\
\text{subject to} & & & f_{L_j}(x, y) = y_{L_j} - B_{L_j} + \sum_t x_{L_j} = 0 \quad (\forall L_j)
\end{align*}
$$

where we have omitted $f_\tau(x, y) = -e^{x_\tau} < 0$ ($\forall \tau$) since it is trivially satisfied.

Consider now the Lagrangian of (P_2),

$$
\Lambda(x, y, (\mu_{L_j})) = f(x, y) + \sum_{L_j} \mu_{L_j} f_{L_j}(x, y).
$$

We remark that KKT conditions (i) and (ii) are immediately satisfied, (i) trivially and (ii) by construction. We need only consider (iii). By the Karush-Kuhn-Tucker theorem, $(x, y)$ is an optimal solution of (P_2) if and only if $\nabla \Lambda(x, y, (\mu_{L_j})) = 0$, with gradient with respect to $(x, y)$, since the subgradient is unique when the function is differentiable.
We have the following derivatives of the Lagrangian $\Lambda$,

\[
\frac{\partial \Lambda}{\partial x_{ru}} = q e^{q x_{ru}} + \sum_{L_j : L_u = ru} \mu_{L_j},
\]

(4.4.2)

\[
\frac{\partial \Lambda}{\partial y_{L_j}} = q e^{q y_{L_j}} + \mu_{L_j}.
\]

(4.4.3)

$(\Rightarrow)$ Assume $(x, y)$ is a local minimum for $(P_2)$. Then the KKT conditions are satisfied, and we can find $\mu$. By (iii) the derivatives (4.4.2), (4.4.3) are zero, so that we obtain

\[
e^{q x_{ru}} = \sum_{L_j : L_u = ru} e^{q y_{L_j}} > 0.
\]

(4.4.4)

A simple calculation yields $d^q_r = \sum_{L : L'^\tau = u} ||\alpha_L||^q_q$.

$(\Leftarrow)$ Assume $d^q_\tau = \sum_{L : L'^\tau = u} ||\alpha_L||^q_q$ for all $\tau$. This defines an optimal solution if we can find $\mu$ such that the KKT conditions are satisfied. But we derived these above. $\square$

### 4.4.1 Proof of Theorem 4.3.5

We show that $\mathbb{P}(\sup ||u|| = c Q_n(\beta_n^0 + \gamma_n u) < Q_n(\beta_n^0)) \geq 1 - \epsilon$, where $\gamma_n = \sqrt{n}(1/\sqrt{n} + \lambda_n), c > 0$ constant. Let $w_L(x) = 1$ if $\varphi_{L_1}(D(\beta_n^0 + xu), 1) = 0, \varphi_{L_1}(D(\beta_n^0 + xu), 1)$ otherwise. Then

\[
\Delta Q = Q_n(\beta_n^0 + \gamma_n u) - Q_n(\beta_n^0)
\]

\[
= \ell_n(\beta_n^0 + \gamma_n u) - \ell(\beta_n^0) - n\lambda_n \sum_{L} \left( \frac{||\beta_n^0 + \gamma_n u||}{w_L(\gamma_n)} - \frac{||\beta_n^0||}{w_L(0)} \right)
\]

\[
= \Delta \ell_n - n\lambda_n \Delta N.
\]

The following argument concerning the log-likelihood follows Fan and Peng (2004).
By third order Taylor series expansion,

\[
\Delta \ell_n = \frac{\gamma_n u^t \nabla_{\beta_n} \ell_n(\beta_n^0) + u^t \nabla_{\beta_n}^2 \ell_n(\beta_n^0) u \gamma_n^2/2 + u^t \nabla_{\beta_n} (u^t \nabla_{\beta_n}^2 (\ell_n(\beta_n^*) u) \gamma_n^3/6}{= I_1 + I_2 + I_3.}
\]

We look at each term individually.

By regularity condition (A2),

\[
|I_1| = |\gamma_n \nabla_{\beta_n} \ell_n(\beta_n^0) u| \\
\leq \gamma_n ||\nabla_{\beta_n} \ell_n(\beta_n^0)|| \cdot ||u|| \\
\ll_p \gamma_n p_n \sqrt{n/p_n} ||u|| \\
\ll \gamma_n^2 n ||u||.
\]

For the second term,

\[
I_2 = \frac{1}{2} u^t \left( \frac{1}{n} \nabla_{\beta_n}^2 \ell_n(\beta_n^0) + I_n(\beta_n^0) \right) u \gamma_n - \frac{1}{2} u^t I_n(\beta_n^0) u \gamma_n^2.
\]

We find by the Chebyshev inequality,

\[
P \left( \left| \frac{1}{n} \nabla_{\beta_n}^2 \ell_n(\beta_n^0) + I_n(\beta_n^0) \right| \geq \frac{\epsilon}{p_n} \right) \leq \frac{p_n^2}{n^2 \epsilon^2} \left( \sum_{i=1}^{p_n} \sum_{j=1}^{p_n} \left( \frac{\partial^2 \ell_n(\beta_n^0)}{\partial \beta_{ni} \partial \beta_{nj}} - \mathbb{E} \left( \frac{\partial^2 \ell_n(\beta_n^0)}{\partial \beta_{ni} \partial \beta_{nj}} \right) \right)^2 \right)
\]

\[
= \frac{p_n^2}{n^2 \epsilon^2} \sum_{i=1}^{p_n} \sum_{j=1}^{p_n} \mathbb{E} \left( \frac{\partial^2 \ell_n(\beta_n^0)}{\partial \beta_{ni} \partial \beta_{nj}} \right)^2 - \mathbb{E}^2 \left( \frac{\partial^2 \ell_n(\beta_n^0)}{\partial \beta_{ni} \partial \beta_{nj}} \right)
\]

\[
< \frac{p_n^2}{n^2 \epsilon^2} \sum_{i=1}^{p_n} \sum_{j=1}^{p_n} c_4
\]

\[
= \frac{p_n^4}{n} \cdot c_4 \ll \frac{p_n^4}{n^2} = o(1)
\]

so that \( I_2 = -\frac{1}{2} u^t I_n(\beta_n^0) u \gamma_n^2 + o_p(1) \).
For the third term,

\[ I_3 = \frac{1}{6} \sum_{L} \sum_{k=1}^{\lvert L \rvert} \sum_{j=1}^{\lvert J \rvert} \sum_{j=1}^{\lvert J \rvert} \frac{\partial^3 l_n(\beta^*)}{\partial \beta_{L_k} \partial \beta_{K_k} \partial \beta_{J_j}} u_{L_i} u_{K_k} u_{J_j} \]

\[ \leq \frac{n}{6} \gamma_n^3 \| u \|^3 \sum_{i=1}^{n} \left( \sum_{L} \sum_{l=1}^{\lvert L \rvert} \sum_{K} \sum_{k=1}^{\lvert K \rvert} \sum_{J} \sum_{j=1}^{\lvert J \rvert} M_{nL_iK_kJ_j}^2 (Y_{ni}, X_{ni}) \right)^{1/2} \]

\[ \ll_p \gamma_n^3 n \| u \|^3 p_n^{3/2} \]

\[ = o_p(n \gamma_n^2 \| u \|^3), \]

since \( p_n^4/n \to 0 \) and \( p_n^2 \lambda_n \to 0 \) by hypothesis. Here \( L, K, J \) run over all lineages.

Thus \( \Delta \ell \ll_p \gamma_n^2 n \| u \| - u^L \ell_n(\beta^0) u n \gamma_n^2/2 + o_p(n \gamma_n^2 \| u \|^3) \). Choosing \( c \) sufficiently large, \( I_2 \) dominates \( I_1 \) uniformly on \( \| u \| = c \), then choosing \( n \) sufficiently large, \( I_2 \) dominates \( I_3 \) uniformly on \( \| u \| = c \).

We next turn to \( \Delta N \). We have \( \Delta N = \sum_L \Delta N_L \) where \( \Delta N_L = \frac{\| \beta_{nL}^0 + \gamma_n u_L \|}{\| w_L(\gamma_n) \|} - \frac{\| \beta_{nL}^0 \|}{\| w_L(0) \|} \).

Suppose \( \beta_{nL}^0 \neq 0 \). Then

\[ |\Delta N_L| = \left| \frac{\| \beta_{nL}^0 + \gamma_n u_L \|}{w_L(\gamma_n)} - \frac{\| \beta_{nL}^0 \|}{w_L(0)} \right| = \left| \frac{\| \beta_{nL}^0 + \gamma_n u_L \| - \| \beta_{nL}^0 \|}{\min\{w_L(\gamma_n), w_L(0)\}} \right|. \]

For \( n \) large enough, \( w_L(\gamma_n) \geq (1 - \xi) w_L(0) \), \( \xi > 0 \) small, so

\[ |\Delta N_L| < \left| \frac{\| \beta_{nL}^0 + \gamma_n u_L \| - \| \beta_{nL}^0 \|}{(1 - \xi) w_L(0)} \right| \ll_p \frac{\gamma_n \| u_L \|}{(1 - \xi) w_L(0)} \ll \gamma_n \| u_L \|. \]

We have

\[ n \lambda_n \sum_L \gamma_n \| u_L \| \leq \sqrt{n} \gamma_n \| u \| \]

where \( L \) runs over all lineages such that \( \beta_{nL}^0 \neq 0 \). From this it follows that \( I_2 \) dominates \( \sum_L |\Delta N_L| \).
Suppose now \( \beta_{0|K} = 0 \) for lineage \( K \). Then
\[
\Delta N_K = \frac{||\beta_{0|K} + \gamma_n u_K||}{w_K(\gamma_n)} - \frac{||\beta_{0|K}||}{w_K(0)} = \frac{\gamma_n||u_K||}{w_K(\gamma_n)} > 0.
\]
Therefore the term \((I_2 - n\lambda_n \sum_K \Delta N_K) < 0\) dominates in \(\Delta Q_n\), where \( K \) runs over all true sparse lineages, \( \beta_{0|K} = 0 \), and we have convergence in probability. Therefore \(||\hat{\beta}_n - \beta_{0|n}|| \ll_p \gamma_n\), which completes the proof of consistency. \(\Box\)

### 4.4.2 Proof of Theorem 4.3.6

We turn now to the sparsity property. Recall that \( \beta_{nL_1} \) is a signed polynomial in \( |\alpha_{nK}|, k = 1, \ldots, |K| \), over lineages \( K \). Consider the polynomial map \( f : \mathbb{R}^n \to \mathbb{R}^n \) given by \( \alpha_n \mapsto \beta_n \), where \( \alpha_n \) is the projection onto the second entry of \( \psi(\beta_n) = (D_n, \alpha_n) \).

When \( \tilde{\alpha}_n \) (equivalently \( \tilde{\beta}_n \)) is non-sparse, the tangent map \( \partial_{\tilde{\alpha}_n} f : T_{\tilde{\alpha}_n}(\mathbb{R}^n) \to T_{\tilde{\beta}_n}(\mathbb{R}^n) \) is non-singular. Since the entries \( \frac{\partial \beta_K}{\partial \alpha_{L_1}} \) are themselves polynomials, they are continuous and hence \( \partial_{\tilde{\alpha}_n} f \) is continuous with respect to \( \alpha_n \) in a neighbourhood of \( \tilde{\alpha}_n \).

We have
\[
\partial_{\alpha_n} f \frac{\partial}{\partial \beta_n} Q_n = \partial_{\alpha_n} f \frac{\partial \ell_n}{\partial \beta_n} - n\lambda_n \partial_{\alpha_n} f \frac{\partial}{\partial \beta_n} \sum_L ||\beta_L|| \frac{1}{w_L} \\
= \partial_{\alpha_n} f \frac{\partial \ell_n}{\partial \beta_n} - n\lambda_n \frac{\partial}{\partial \alpha_n} ||\alpha_n|| \\
= \partial_{\alpha_n} f \frac{\partial \ell_n}{\partial \beta_n} - n\lambda_n \text{sign}(\alpha_n)
\]
where \( \text{sign}(\alpha_n) \) is understood as the vector of signs \( (\text{sign}(\alpha_{n1}), \ldots, \text{sign}(\alpha_{np_n})) \).

Assume hereon that \( ||\beta_n - \beta_{n}|| \ll_p \sqrt{p_n/n} \).
By second order Taylor expansion, we have

\[
\frac{\partial \ell_n}{\partial \beta_{L_l}} = \frac{\partial \ell_n}{\partial \beta_{L_l}} \bigg|_{\beta_0^n} + \sum_{K} \sum_{k=1}^{\mid K \mid} \frac{\partial^2 \ell_n}{\partial \beta_{K_k} \partial \beta_{L_l}} \bigg|_{\beta_0^n} (\beta_{nK_k} - \beta_{nK_k}^0) \\
+ \frac{1}{2} \sum_{j=1}^{\mid J \mid} \sum_{K} \sum_{k=1}^{\mid K \mid} \frac{\partial^3 \ell_n}{\partial \beta_{K_k} \partial \beta_{J_j} \partial \beta_{L_l}} \bigg|_{\beta_0^n} (\beta_{nJ_j} - \beta_{nJ_j}^0) (\beta_{nK_k} - \beta_{nK_k}^0) \\
= G_1 + G_2 + G_3
\]

where \( J, K \) run over all lineages and \( \beta_n^* \) lies between \( \beta_n^0 \) and \( \beta_n \).

We have from the proof of consistency that \( G_1 \ll_p \sqrt{mp_n} \). Bounds for \( G_2 \) and \( G_3 \) are provided in Fan and Peng (2004), which we derive here for completeness. For \( G_2 \), we have that

\[
G_2 = \sum_{j=1}^{\mid J \mid} \sum_{j=1}^{\mid J \mid} \left[ \frac{\partial^2 \ell_n}{\partial \beta_{nJ_j} \partial \beta_{nK_k}} \bigg|_{\beta_n^0} - \mathbb{E} \left( \frac{\partial^2 \ell_n}{\partial \beta_{nJ_j} \partial \beta_{nK_k}} \bigg|_{\beta_n^0} \right) \right] \cdot (\hat{\beta}_{nJ_j} - \beta_{nJ_j}^0) \\
+ \sum_{j=1}^{\mid J \mid} \sum_{j=1}^{\mid J \mid} \mathbb{E} \left( \frac{\partial^2 \ell_n}{\partial \beta_{nJ_j} \partial \beta_{nK_k}} \bigg|_{\beta_n^0} \right) (\hat{\beta}_{nJ_j} - \beta_{nJ_j}^0) \\
= F_1 + F_2
\]

By Cauchy-Schwarz inequality and \( ||\hat{\beta}_n - \beta_n^0|| \approx_p \sqrt{p_n/n} \),

\[
|F_2| = \left| n \sum_{j=1}^{\mid J \mid} \sum_{j=1}^{\mid J \mid} \mathcal{I}_{nJ_jK_k}(\beta_n^0)(\hat{\beta}_{nJ_j} - \beta_{nJ_j}^0) \right| \ll_p n(\sqrt{p_n/n}) \left( \sum_{j=1}^{\mid J \mid} \sum_{j=1}^{\mid J \mid} \mathcal{I}_{nJ_jK_k}^2(\beta_n^0) \right)^{1/2}.
\]

By the regularity condition (A2) on the eigenvalues of \( \mathcal{I}_n \), we have \( \sum_j \sum_{j=1}^{\mid J \mid} \mathcal{I}_{nJ_jK_k}^2 \ll_p 1 \), so that \( |F_2| \ll_p \sqrt{mp_n} \). For \( F_1 \), by the Cauchy-Schwarz inequality, we have

\[
|F_1| \leq ||\hat{\beta}_n - \beta_n^0|| \left[ \sum_{j=1}^{\mid J \mid} \sum_{j=1}^{\mid J \mid} \left( \frac{\partial^2 \ell_n}{\partial \beta_{nJ_j} \partial \beta_{nK_k}} \bigg|_{\beta_n^0} - \mathbb{E} \left( \frac{\partial^2 \ell_n}{\partial \beta_{nJ_j} \partial \beta_{nK_k}} \bigg|_{\beta_n^0} \right) \right)^2 \right]^{1/2}
\]
so that by regularity condition (A2), \(|F_1| \ll_p \sqrt{p/n} \sqrt{np_n} \ll \sqrt{np_n} \). Thus \(|G_2| \ll_p \sqrt{np_n} \).

For \(G_3\),

\[
2G_3 = \sum_{J} \sum_{H} \frac{\partial^3 \ell_n(\beta_n^*)}{\partial H_i \partial J_j \partial K_k} (\hat{\beta}_{nJ} - \beta_{nJ}^0)(\hat{\beta}_{nH_i} - \beta_{nH_i}^0)
\]

\[
= \sum_{J} \sum_{H} \frac{\partial^3 \ell_n(\beta_n^*)}{\partial H_i \partial J_j \partial K_k} \left( \hat{\beta}_{nJ} - \beta_{nJ}^0 \right) \left( \hat{\beta}_{nH_i} - \beta_{nH_i}^0 \right)
\]

\[
+ \sum_{J} \sum_{H} \sum_{h} \mathbb{E} \left\{ \frac{\partial^3 \ell_n(\beta_n^*)}{\partial H_i \partial J_j \partial K_k} \right\} (\hat{\beta}_{nJ} - \beta_{nJ}^0)(\hat{\beta}_{nH_i} - \beta_{nH_i}^0)
\]

\[= F_3 + F_4.\]

By Cauchy-Schwarz inequality and the regularity conditions on \(\ell_n\),

\[
F_3^2 \leq \sum_{J} \sum_{H} \sum_{h} \left[ \frac{\partial^3 \ell_n(\beta_n^*)}{\partial H_i \partial J_j \partial K_k} \right]^2 \mathbb{E} \left\{ \frac{\partial^3 \ell_n(\beta_n^*)}{\partial H_i \partial J_j \partial K_k} \right\} \left( \hat{\beta}_n - \beta_n^0 \right)^4
\]

\[\ll_p \left( \frac{p^2}{n^2} \right) \cdot (n^2) = o_p(np_n)\]

and

\[|F_4| \leq C_5^{1/2} np_n \| \hat{\beta}_n - \beta_n^0 \|^2 \ll_p p_n^2 = o_p(\sqrt{np_n}),\]

hence \(G_3 \ll_p \sqrt{np_n} \) and \(G_1 + G_2 + G_3 \ll_p \sqrt{np_n} \).

Consider now the elements of the tangent map \(\partial \alpha_n f \cdot \frac{\partial \beta_{nL_i}}{\partial \alpha_n K_k} \). We may explicitly write out \(\beta_{nL_i}\) in terms of \(\alpha_n\),

\[\beta_{nL_i} = \alpha_{nL_i} \cdot \prod_{t=1}^{T} \sum_{K: K_t = L_t} \| \alpha_K \|. \quad (4.4.6)\]
From equation (4.4.6), we obtain

$$\frac{\partial \beta_{nL_i}}{\partial \alpha_{nK_k}} = \begin{cases} 
0 & \text{if } L' \neq K' \forall t \\
\alpha_{nL_i} \cdot \text{sign}(\alpha_{nL_i}) \cdot \sum_{t=1}^{T} \prod_{u=1, u \neq t}^{T} \sum_{J: J' = L_t} ||\alpha_{nJ}|| \\
+ \text{sign}(\alpha_{nL_i}) \cdot \prod_{t=1}^{T} \sum_{J: J' = L_t} ||\alpha_{nJ}|| & \text{if } L_t = J_j \\
\alpha_{nL_i} \cdot \text{sign}(\alpha_{nK_k}) \cdot \sum_{t \in S} \prod_{u=1, u \neq t}^{T} \sum_{J: J' = L_u} ||\alpha_{nJ}|| & \text{otherwise, with} \\
S = \{t : L_t = K^t\}. 
\end{cases}$$

Suppose that $\beta^0_{n\tau^t} = 0$ for some taxon $\tau^t$ of taxon level $t$. Then on inspection for $L_t = \tau^t$, we find $\frac{\partial \beta_{nL_t}}{\partial \alpha_{nK_k}} \ll_p \left(\frac{p_n}{n}\right)^{1/(2T+2)}$. From $(\partial_{\alpha_n} f)^{-1}(\partial_{\alpha_n} f) = I_{p_n}$, we obtain the relationship

$$\sum_{j=1}^{p_n} \frac{\partial \beta_{nj}}{\partial \alpha_{nj}} \cdot \frac{\partial \alpha_{nj}}{\partial \beta_{nj}} \ll_p \left(\frac{p_n}{n}\right)^{1/(2T+2)} \sum_{j=1}^{p_n} \frac{\partial \alpha_{nj}}{\partial \beta_{nj}},$$

so that $(n/p_n)^{1/(2T+2)} \ll_p \sum_{j=1}^{p_n} \frac{\partial \alpha_{nj}}{\partial \beta_{nj}}$.

Without loss of generality, assume $\alpha_n > 0$, since we can always transfer the sign from the coefficients to the covariates at every $n$. Evaluated at our $\sqrt{n/p_n}$-consistent estimate $\hat{\beta}_n$, we have

$$0 = (\partial_{\alpha_n} f)^{-1}(\partial_{\alpha_n} f) \nabla_{\beta_n} \ell_n(\hat{\beta}_n) - n\lambda_n (\partial_{\alpha_n} f)^{-1} \text{sign}(\alpha_n)$$

$$= \nabla_{\beta_n} \ell_n(\hat{\beta}_n) - n\lambda_n (\partial_{\alpha_n} f)^{-1} \text{sign}(\alpha_n).$$

Thus for a sparse taxon $L_t \in \tau$,

$$\nabla_{\beta_n} \ell_n(\hat{\beta}_n) = n\lambda_n (\partial_{\alpha_n} f)^{-1} \text{sign}(\alpha_n)$$

$$n\lambda_n \left(\frac{n}{p_n}\right)^{1/(2T+2)} \ll_p n\lambda_n \sum_{j=1}^{p_n} \frac{\partial \alpha_{nj}}{\partial \beta_{nj}} \ll_p \sqrt{np_n}$$

and hence $n \ll_p p_n^{T+2}$, a contradiction. Therefore $\lim_{n \to \infty} P(\delta(\hat{\beta}_{n\tau}) = \delta(\beta^0_{n\tau})) = 1$ for all taxa $\tau$. □
4.4.3 Proof of Theorem 4.3.7

(i) The proof of model consistency is immediate from Theorem 4.3.6.

(ii) Let $A_n \subseteq \{1, \ldots, p_n\}$ be the subset of indices for nonzero parameters. We have that there exists a $\sqrt{n/p_n}$ maximizer $\hat{\beta}_n = (\hat{\beta}_{nA_n}, 0)$ of $Q_n$. We take to writing $Q_n(\beta_{nA_n}) = Q_n(\hat{\beta}_{nA_n}, 0)$ through abuse of notation. The following argument concerning the log-likelihood $\ell_n$ follows Zhou and Zhu (2010).

By Taylor expansion of $\nabla_{\beta_n} Q_n$ about $0_{nA_n}$, we find

$$
\frac{1}{n} \left( \nabla_{\beta_n}^2 \ell_n(\beta_{nA_n}^0)(\hat{\beta}_{nA_n} - \beta_{nA_n}^0) - \nabla_{\beta_n} P_{\lambda}(\hat{\beta}_{nA_n}) \right)
$$

$$
= -\frac{1}{n} \left( \nabla_{\beta_n} \ell_n(\beta_{nA_n}) + \frac{1}{2}(\hat{\beta}_{nA_n} - \beta_{nA_n}^0)^T \nabla_{\beta_n}^2 \ell_n(\beta_{nA_n}^*) \nabla_{\beta_n} \ell_n(\beta_{nA_n}^*)(\hat{\beta}_{nA_n} - \beta_{nA_n}^0) \right)
$$

from the relationship $0 = \nabla_{\beta_n} Q_n(\hat{\beta}_n)$.

By Cauchy-Schwarz inequality,

$$
\left\| \frac{1}{2n} (\hat{\beta}_{nA_n} - \beta_{nA_n}^0)^T \nabla_{\beta_n}^2 \ell_n(\beta_{nA_n})(\hat{\beta}_{nA_n} - \beta_{nA_n}^0) \right\|^2
$$

$$
\leq \frac{1}{4n^2} \sum_{i=1}^n ||\hat{\beta}_{nA_n} - \beta_{nA_n}^0||^4 \sum_{j_1 \in A_n} \sum_{j_2 \in A_n} \sum_{j_3 \in A_n} M_{nj_1,j_2,j_3}^3 (X_{ni}, Y_{ni})
$$

$$
\ll_p p_n^5/n^2 = o_p(1/n)
$$

By Lemma 8 in Fan and Peng (2004), obtain

$$
\left\| \left( \frac{1}{n} \nabla_{\beta_n}^2 \ell_n(\beta_{nA_n}^0) + I_n(\beta_{nA_n}^0) \right) (\hat{\beta}_{nA_n} - \beta_{nA_n}^0) \right\| = o_p(1/\sqrt{n}).
$$

We turn now to the penalty $p_{\lambda}$. Since $||\beta_n||$ is a polynomial in $\alpha_n$ of degree greater than 1, we have $\frac{1}{n} \nabla_{\beta_n} ||\alpha_n|| = o_p(\frac{1}{n} ||\nabla_{\beta_n}||\beta_n||||) = o_p(p_n/n)$. 
Therefore from the Taylor expansion of $\nabla_{\beta_n} Q$, we have

$$I_n(\beta^0_{n,A_n})(\hat{\beta}_{n,A_n} - \beta^0_{n,A_n}) + o_p(p_n/n) = \frac{1}{n}\nabla_{\beta_n} \ell_n(\beta^0_{n,A_n}) + o_p(1/\sqrt{n}).$$

Following Fan and Peng (2004),

$$\sqrt{n}A_n I_n^{-1/2}(\beta^0_{n,A_n})\left(I_n(\beta^0_{n,A_n})(\hat{\beta}_{n,A_n} - \beta_{n,A_n}) + o_p(p_n/n)\right)$$

$$= \sqrt{n}A_n I_n^{-1/2}(\beta^0_{n,A_n})\left(\frac{1}{n}\nabla_{\beta_n} \ell_n(\beta^0_{n,A_n}) + o_p(1/\sqrt{n})\right)$$

$$\sqrt{n}A_n I_n^{1/2}(\beta^0_{n,A_n})(\hat{\beta}_{n,A_n} - \beta_{n,A_n}) \sim \sqrt{n}A_n I_n^{-1/2}(\beta^0_{n,A_n})\frac{1}{n}\nabla_{\beta_n} \ell_n(\beta^0_{n,A_n})$$

$$\sim N(0, \Sigma)$$

where $A_n$ and $\Sigma$ are as described in the statement of the theorem. □
Clostridium difficile: A public health concern

5.1 Background

Clostridium difficile (C. difficile) infection (CDI) is the most frequent cause of healthcare-associated infections and its rates are growing in the community, Loo et al. (2005); Brandt et al. (2012). One of the major risk factors for developing CDI is through antibiotics. The healthy and diverse commensal bacteria residing within the colon are the major defense against the growth of C. difficile. Antibiotics kill these bacteria and allow C. difficile to colonize and multiply, produce toxins, and cause disease. The available antibiotic treatments for this infection are metronidazole, vancomycin, and fidaxomicin. The efficacy of these antibiotics is limited as vancomycin and metronidazole also suppress the growth of anaerobic bacteria such as the Bacteroides fragilis group which protect against C. difficile proliferation, Hopkins and Macfarlane (2002). The persistent disruption of healthy colonic flora may in part explain the reason for recurrences following a course of treatment with these antibiotics.
An alternative to antibiotic therapy, in particular for recurrent and refractory CDI, infuses healthy gut bacteria directly into the colon of infected patients to outcompete *C. difficile* by a procedure known as fecal microbiota transplantation (FMT). FMT is a process in which a healthy donor’s stool is collected and the bacterial component is administered to an infected patient. This can be performed using a colonoscope, nasogastric tube, or enema. FMT serves to reconstitute the altered colonic flora, in contrast to treatment with antibiotic(s), which can further disrupt the establishment of key microbes essential in preventing recurrent CDI. The literature reveals a cumulative clinical success rate of 92% in confirmed recurrent CDI cases, Gough et al. (2011).

There has been a growing interest in the microbiome of CDI patients, Manges et al. (2010); Vincent et al. (2013), especially those involved with FMT, Hamilton et al. (2013); Song et al. (2013a); Weingarden et al. (2014). In case of the latter, there are differences in the administration route; donor selection criteria, some used family members, while others used a pool of donors; sample sizes, although these were small in all studies; and sequencing procedures and equipment. Despite these differences, there are two central points of agreement across all studies. The first is that CDI patients have low diversity in their microbiome, and that after receiving an FMT(s), their microbiome diversity increased. The second is that CDI patients who were successfully treated with FMT undergo changes in their microbiome that, at least initially, have structural similarities to that of their donors, Shahinas et al. (2012).

The report by Schubert et al. (2014) develops logistic regression models incorporating enterotype with *Clostridium* removed in order to predict diarrheal state and cause. Briefly, they clustered all the samples into 13 community types based on similarity in OTU
composition, and used these as a factor in their model. Their finding that *C. difficile*-associated diarrhea was matched with several distinct enterotype clusters was consistent with some of the variance we observed in the pre-FMT communities. They found that *C. difficile*-associated diarrhea cases have, relative to healthy enterotypes, decreased *Bacteroides*, *Ruminococcaceae*, *Lachnospiraceae* and some *Lachnospiraceae* species. This is the only paper to date that we are aware of that directly incorporates metagenomic covariates into a logistic regression framework for examining CDI. Indeed, the question as to what bacteria in patient and donor colonic microbial systems play crucial roles in helping or hindering clinical resolution of CDI is an important medical question. It is also important to develop a statistical methodology attuned to the idiosyncratic nature of the problem.

5.2 Application

5.2.1 Description

We will examine the microbiome data coming from a subset of the CDI-patients treated with FMT in Lee et al. (2014), covering the period 2008–2012. All *C. difficile* infections were confirmed by in-hospital, real-time, polymerase chain reaction (PCR) testing for the toxin B gene. This study sequenced the forward V3-V5 region of the 16S rRNA gene from 17 CDI patients who were treated with FMT(s). A pre-FMT, and the corresponding post-FMT were sequenced. All sequencing was performed on the 454 Life Sciences, GS Junior Titanium Series. The E.Z.N.A. Stool DNA Kit (Omega BIO-TEK, Norcross, Georgia) was used to extract the DNA from the fecal samples following the ‘stool DNA protocol for pathogen detection’. Subsequent DNA amplification was done using PCR forward and re-
verse primers. One round of DNA amplicon purification was performed using the QIAquick PCR Purification Kit (Qiagen, Valencia, CA) followed by two rounds of purification using Agencourt AMPure XP beads (Beckman Coulter Inc., Mississauga, ON). The final amplification was performed and the sequencing subsequently performed. The bioinformatics software mothur was used as the primary means of processing and quality-filtering reads and calculating statistical indices of community structure; see Schloss et al. (2009) for a breakdown of the mothur processing pipeline. The mothur standard operating procedure for pyrosequencing was followed to obtain OTUs. The full details are recorded in Pinder (2013).

From the 17 selected patients, we note that 13 of these patients responded to a single FMT. We used a taxonomy describing phylum, class, order, family, and genus. The predictors consisted of the relative abundances of 220 pre-FMT OTUs and 347 post-FMT OTUs. Here we consider logistic regression to model the response for two scenarios. First we wish to know whether the pre-FMT microbiome could predict a clinical response to an FMT. Second, we wish to know whether the composition of the post-FMT microbiome could anticipate the need for additional FMTs. To select the tuning parameter for the $\Phi$-LASSO, we perform leave-one-out cross-validation (LOO-CV). The optimal tuning parameter was selected by AUC (area under the curve) and BS (Brier score). This is the same dataset where phylum interaction was investigated, Martinez et al. (2016).

**Definition 5.2.1.1** Let $y_i, i = 1, ..., n$ be a set of responses observed from $\{0, 1\}$ and let $p_i, i = 1, ..., n$ be the set of corresponding predicted probabilities for each event. Then the
**Brier score** is defined as
\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2.
\]

**Definition 5.2.1.2** Let \( y_i, i = 1, ..., n \) be a set of responses observed from \( \{0, 1\} \) and let \( p_i, i = 1, ..., n \) be the set of corresponding predicted probabilities for each event. Further suppose that \( y_i = 0 \) for \( i = 1, ..., n_0 \) and \( y_i = 1 \) for \( i = n_0 + 1, ..., n \). Then the **area under the curve** is defined as
\[
\frac{1}{n_0(n - n_0)} \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n} I(p_j \geq p_i),
\]
where \( I \) is the indicator function taking the value 1 when the event is true, 0 otherwise.

### 5.2.2 Analysis

We are challenged by a sparse predictor matrix as well as a small sample size. Since AUC tends to assign a perfect score to the null model by the way it handles ties, we incorporate BS to compensate for this affect. The results of logistic regression using the pre-FMT OTUs as covariates are captured in Figure 5.1. Using the globally optimal BS of 0.208, the corresponding tuning parameter is \( \lambda = 0.0067 \). Relative to the BS, the locally optimal AUC value is 0.846, with the corresponding tuning parameter of \( \lambda = 0.0049 \). The globally optimal AUC value is 0.865 which would be associated with the null model. Below the figure in Table 5.1, we display the family and genus of the selected OTUs. Using LOO-CV, the frequency along with the averaged estimate and LOO-CV standard errors are reported. Due to the small sample size the variability in the parameter estimates is large which is to be expected. Concentrating on the frequency using LOO-CV we notice that OTU 7 associated with the genus *Lactobacillus* seems to have some positive predictability.
This is consistent with some earlier findings from a different FMT patient cohort, Shahinias et al. (2012).

For the post-FMT OTUs, we obtain a globally optimal BS of 0.386, $\lambda = 0.0056$, with corresponding AUC value 0.923, $\lambda = 0.0065$. The globally optimal AUC is 0.961. This is displayed in Figure 5.2. In Table 5.2, we display the family and genus of the selected OTUs. Using LOO-CV, the frequency along with the averaged estimate and LOO-CV standard errors are reported. Again due to the small sample size the variability in the parameter estimates is large. Concentrating on the frequency using LOO-CV we notice that OTU 5 associated with the family *Enterococcaceae* and the corresponding genera *Enterococcus*, as well as OTU 17 corresponding to the family *Bacteroidaceae* and genera *Bacteroides* appear to have some positive predictability. This again is consistent with some earlier results, Shahinias et al. (2012).

We highlight above some of the more obvious interpretations that is consistent with what has been observed in other studies. Indeed, results of this work will interest researchers and companies working toward refinement of FMT, *via* establishment of central stool banks or creation of synthetic stool. With results obtained using larger sample sizes it will be possible to select species that come from the genera identification and culture them in a laboratory biochemistry setting. Thus one could strategically pool samples to achieve a desired composition, to supplement stool with specific microorganisms or to be able to prepare recipient-specific synthetic stool as *per* the RePOOPulate project, Petrof et al. (2013). At this point however, we will end this discussion with the comment that a more in depth metageonomic analysis of a recent clinical trial, Lee et al. (2016), is under investigation.
Figure 5.1: AUC and BS obtained by leave-one-out cross-validation for pre-FMT OTUs. Labeled are the tuning parameters by (a) AUC ($\lambda = 0.0049$) and (b) BS ($\lambda = 0.0067$).

```
<table>
<thead>
<tr>
<th>OTU</th>
<th>Family</th>
<th>Phylogeny</th>
<th>Genus</th>
<th>Frequency $\hat{\beta}$ (SE)</th>
<th>$\lambda = 0.0049$</th>
<th>Frequency $\hat{\beta}$ (SE)</th>
<th>$\lambda = 0.0067$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Enterobacteriaceae</td>
<td>Klebsiella</td>
<td>0.24 10.8 (41.6) 0.35 -4.65 (13.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Enterobacteriaceae</td>
<td>Escherichia/Shigella</td>
<td>0.59 -49.4 (122) 0.88 -107 (143)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Streptococcaceae</td>
<td>Streptococcus</td>
<td>0.82 -304 (323) 0.76 -395 (425)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Lachnospiraceae</td>
<td>Blautia</td>
<td>0.82 239 (281) 0.18 20.4 (66.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Enterococcaceae</td>
<td>Enterococcus</td>
<td>0.47 -94.9 (137) 0.65 -164 (264)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Lactobacillaceae</td>
<td>Lactobacillus</td>
<td>0.53 -15.5 (29.2) 0.82 -29.8 (29.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Lactobacillaceae</td>
<td>Lactobacillus</td>
<td>0.88 349 (335) 0.88 387 (415)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Lachnospiraceae</td>
<td>unclassified</td>
<td>0.53 -60.9 (77.5) 0.59 -68.6 (85.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
<td>0.29 78.5 (127) - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Veillonellaceae</td>
<td>unclassified</td>
<td>0.65 -108 (141) - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Veillonellaceae</td>
<td>unclassified</td>
<td>- - 0.47 -97.9 (142)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
<td>0.76 176 (258) - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
<td>- - 0.53 55.6 (107)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Table 5.1: The selected OTUs for pre-FMT microbiomes at the tuning parameters selected by AUC ($\lambda = 0.0049$) and BS ($\lambda = 0.0067$) leave-one-out cross-validation on relative abundances.
Figure 5.2: AUC and BS obtained by leave-one-out cross-validation for post-FMT OTUs. Labeled are the tuning parameters by (a) AUC ($\lambda = 0.0065$) and (b) BS ($\lambda = 0.0056$).

Table 5.2: The selected OTUs for post-FMT microbiomes at the tuning parameters selected by AUC ($\lambda = 0.0065$) and BS ($\lambda = 0.0056$) leave-one-out cross-validation on relative abundances.
Conclusion

We have presented a novel method of predictive modelling with microbial systems with emphasis on application to microbial therapeutics. The work is incomplete in an essential way, the very low sample size. Ordinarily, the data would be split into at least three independent sets, (i) the first to be used in selecting tuning parameters via cross-validation, (ii) the second to be used in fitting the parameters, and (iii) the third for evaluating predictive performance. Performing all three on the same data results in over-optimistic over-fitting. However, with a sample size of \( n = 17 \), we are restricted to this situation. Indeed, the sample size for the dataset in Chapter 5 is insufficient for cross-validation. A larger data set associated with the recent study Lee et al. (2016), counting between 150 and 200 data points, is currently being assembled, and it is on this expanded data set that a full analysis may be completed.

We have mentioned limits to our resolution due to the truncated 16S window. However, sequencing technology is always improving. Technology sufficient to sequence the 16S+23S rRNA region currently exists, and work is being performed to bring this to
massively parallel scales, Bowman et al. (2015). When this day arrives, our method will be prepared to handle it.

On the theoretical side, there is more work to be performed. Since use of a penalty results in \textit{maximum a posteriori} estimation, the method is suitable for Bayesian analysis. For the LASSO, the associated prior is the multivariate Laplacian on $\beta$ with covariance zero between parameters. For the $\Phi$-LASSO, the associated prior is multivariate Laplacian on $\alpha$, with the covariance structure determined by the relationships encoded in the taxonomy. We have performed simulations from a frequentist perspective, however these may be performed under a Bayesian framework. This would further facilitate sensitivity analysis for the method. For this to be practical, a more efficient implementation of the $\Phi$-LASSO needs to be developed.

An issue not explicitly addressed in our analysis is attenuation bias. We approximate the theoretical true species relative abundance matrix $X_s$ with the matrix of OTUs $X$. It has been observed that such errors-in-the-covariates shrinks parameters towards zero. A systematic and predominantly laboratory based analysis of error propagation through the collection and sequencing process will be necessary for a complete treatment.

We have focused in this thesis on predicting optimal composition of donor microbiome for successful FMT. The $\Phi$-LASSO is not restricted to this. Even now we are expanding its application to generalized mixed models, survival analysis, to metabolomics, and to study of the gut-brain axis.

One of the examiners inquired about using the $\Phi$-LASSO for ultra-high dimensional problems, and in particular when there are too many weak signals. As rightly pointed out, many small effects go into determining an individual’s adult height. There are two
contrasting perspectives in this field. For example Hastie et al. (2001) promote their “bet on sparsity” principle, whereas Burnham and Anderson (2002) take the perspective that all models, and in particular in biological applications, are ultimately infinite dimensional and this should be taken into account. The perspective to take requires deep reflection on the nature of the problem being asked. The oracular version of the $\Phi$-LASSO employs an $l_{1/2}$-penalty at the individual parameter level. Due to the curvature of the $l_{1/2}$-norm, there is a relatively much higher penalty on weaker signals relative to stronger. Compare this to the soft-thresholded penalties like SCAD, where the penalties for weak signals grows linearly. Thus we might expect a tendency for the $\Phi$-LASSO to omit weak signals, unless there are also strong signals in their lineage. One possible modification is to replace the $l_{1/2}$-penalty with the SCAD. One consequence of this is that the group coefficients will become bounded, conditioned on the tuning parameters. Alternatively, we may choose to sacrifice model consistency and use the original $\Phi$-LASSO. In this scenario, the combined mass of weak signals would support each other, due to the relationship $d_k^t = ||\alpha_k^t||_1$, so that if a lineage consists of many weak signals, it would still be selected. I thank the examiner for his excellent question, and look forward to pursuing these questions theoretically and empirically.
Bibliography

Acinas, S., L. Marcelino, V. Klepac-Ceraj, and M. Polz
2004. Divergence and redundancy of 16s rRNA sequences in genomes with multiple rrn

Bäckhed, F., H. Ding, T. Wang, L. Hooper, G. Koh, A. Nagy, C. Semenkovich, and J. Gordon
2004. The gut microbiota as an environmental factor that regulates fat storage. PNAS,

Bakin, S.
Australian National University, Canberra.

Bansal, A. and T. Meyer

Berger, J.
1985. Statistical decision theory and Bayesian analysis, 2 edition. New York: Springer-
Verlag.

Bhattacharya, R. and V. Patrangenaru
2014. Statistics on manifolds and landmarks based image analysis: A nonparametric
theory with applications,. JSPI, 145:1–22.

Bien, J., J. Taylor, and R. Tibshirani

Bobrovsky, B., E. Mayer-Wolf, and M. Zakai

Bowman, B., M. Kin, Y-J. Cho, and J. Korlach
2015. Long-read, single molecule, real-time (SMRT) DNA sequencing for metagenomic
applications. In Metagenomics for microbiology, J. Izard and M. Rivera, eds., chapter 2,


Colwell, R.  


Fan, J. and J. Lv

Fan, J. and H. Peng

Fan, J., L. Xue, and H. Zou

Fierer, N., M. Hamady, C. Lauber, and R. Knight

Fiers, W., R. Contrera, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. van den Berghe, G. Volckaert, and M. Ysebaert


Frank, I. and J. Friedman

Fredericks, D. and D. Relman

Friedman, J., T. Hastie, and R. Tobshirani

Fu, W. and K. Knight
Garcia, T., S. M"uller, R. Carroll, and R. Walzern
2013. Identification of regressor groups, subgroups, and individuals via regularization

Garthwaite, P., I. Jolliffe, and B. Jones

Gough, E., H. Shaikh, and A. Manges
2011. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy)

Hall, P., J. Marron, and A. Neeman
2005. Geometric representation of high dimension, low sample size data. *JRSS B*,
67(3):427–44.

Hamilton, M., A. Weingarden, T. Unno, A. Khoruts, and M. Sadowsky
2013. High-throughput DNA sequence analysis reveals stable engraftment of gut
microbiota following transplantation of previously frozen fecal bacteria. *Gut Microbes*,
4(2):125–133.

Harrington, C., A. Del Casale, J. Kennedy, H. Neve, B. Picton, M. Mooij, F. O’Gara,
L. Kulakov, M. Larkin, and A. Dobson
2012. Evidence of bacteriophage-mediated horizontal transfer of bacterial 16S rRNA
genes in the viral metagenome of the marine sponge Hymeniacidon. *Microbiology*,
158:2789–95.

Harth, E., J. Romero, R. Torres, and R. Espejo
2007. Intragenomic heterogeneity and intergenomic recombination among Vibrio para-

Hastie, T., R. Tibshirani, and J. Friedman
York: Springer.

Hoerl, A.
1962. Application of ridge analysis to regression problems. *Chemical Engineering
Progress*, 58:54–9.

Hoerl, A. and R. Kennard

Hopkins, M. and G. Macfarlane
2002. Changes in predominant bacterial populations in human faces with age and with
Horgan, R. and L. Kenny

Ioannidis, J.

Jenatton, R., J-Y. Audibert, and F. Bach

Kämpfer, P. and S. Glaeser

Kenny, E., N. Timpson, M. Sikora, M.-C. Yee, A. Morenon-Estrada, C. Eng, S. Huntsman, E. Burchard, M. Stoneking, C. Bustamante, and S. Myles


Kim, P., S. Pinder, and S. Rush

Kim, S. and E. Xing

Kim, Y., H. Choi, and H-S. Oh

Kitahara, K. and K. Miyazaki

Knight, K. and W. Fu
Konstantinidis, K. and E. Stackebrandt

Koonin, E., K. Makarova, and L. Aravind

Kozich, J., S. Westcott, N. Baxter, S. Highlander, and P. Schloss

Kunin, V., A. Copeland, A. Lapidus, K. Mavromantis, and P. Hugenholtz

Kwiatkowski, D.

Le Cam, L.


Leeb, H. and B. Pötscher

Leeb, H. and B. Pötscher
Leeb, H. and B. Pötscher


Lehman, E. and G. Casella

Lenski, R.

Ley, R., F. Bäckhed, P. Turnbaugh, C. Lozupone, R. Knight, and J. Gordon

Ley, R., P. Turnbaugh, S. Klein, and J. Gordon

Lim, M. and T. Hastie

Liu, J. and J. Ye

Lockhart, R., J. Taylor, R. Tibshirani, and R. Tibshirani


Lv, J. and Y. Fan

Manges, A., A. Labbe, V. Loo, J. Atherton, M. Behr, L. Masson, P. Tellis, and R. Brousseau
Martinez, D., C. Lee, P. Kim, and W. Mio

McCullagh, P.

McCullagh, P. and J. Nelder


McKinney, M.

McMurdie, P. and S. Holmes

Meier, L., S. van de Geer, and P. Bühlmann

Mizrahi-Man, O., E. Davenport, and Y. Gilad

Nelder, J. and R. Wedderburn

O’Connor, J., S. Johnson, and D. Gerding

Ogunseitan, O.

Park, T. and G. Casella

Petrof, E., G. Gloor, S. Vanner, S. Weese, D. Carter, and M. Daigneault

Pinder, S.

Pötscher, B.

Poutanen, S. and A. Simor

R Core Team

Rainey, F., B. Hollen, and A. Small

Raskutti, G., M. Wainwright, and B. Yu

Rivers, T.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J-C. Sanchez, and M. Mller
2011. proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics, 12:77.

Rockafellar, R.

Roth, V. and B. Fischer

Rush, S., S. Pinder, M. Costa, and P. Kim


Sprent, J. and P. Sprent

Stackebrandt, E. and B. Goebel

Stanier, R. and C. van Niel

Stewart, P.

Surette, M. and B. Bassler

Tibshirani, R.

Tringe, S. and E. Rubin

Turnbaugh, P., R. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. Gordon

Turnbaugh, P., R. Ley, M. Mahowald, V. Magrini, E. Mardis, and J. Gordon


van de Geer, S. and P. Bühlmann

van der Vaart, A.


Woese, C. and G. Fox

Wu, T. and K. Lange

Yap, W., Z. Zhang, and Y. Wang

Ye, F. and C-H. Zhang

Yuan, M. and Y. Lin

Yuan, M. and H. Zou

Yukawa, M. and S-I. Amari

Zhang, C-H.

Zhang, C-H. and T. Zhang


Zhang, T.
Zhao, P., G. Rocha, and B. Yu

Zhou, N. and J. Zhu

Zinder, S. and M. Dworkin

Zou, H.

Zou, H. and T. Hastie

Zou, H. and R. Li

Zou, H. and H. Zhang

Zuckerbrandt, E. and L. Pauling
Appendix A

Φ-LASSO source code

A.0.1 R-code

```r
# File: ophyglm.r
# Versions: 2.0
# Date: 2016-06-17
# Author: Stephen Rush <srush01@uoguelph.ca>
# Maintainer: Stephen Rush <srush01@uoguelph.ca>
# Description: Regularized regression for variable selection using
taxonomic information in the sparsity-inducing prior.
# This is oracular version of phy-lasso, which effectively
# selects important lineages and important members of these
# lineages.

library("glmnet")
dyn.load("~/Documents/Phylasso/Phylasso code/Fortran/redistribute6.so")

phyglm <- function (x, y, taxonomy, family = c("gaussian", "binomial",
"poisson", "multinomial", "cox", "mgaussian"), weights,
offset = NULL, alpha = 1, nlambda = 100,
lambda.min.ratio = ifelse(nobs < nvars, 0.01, 1e-08), lambda = NULL,
standardize = FALSE, intercept = TRUE, thresh = 1e-06,
dfmax = nvars + 1, pmax = min(dfmax * 2 + 20, nvars), exclude,
penalty.factor = rep(1, nvars), lower.limits = -Inf,
upper.limits = Inf, maxit = 1e+05,
type.gaussian = ifelse(nvars < 500, "covariance", "naive"),
type.logistic = c("Newton", "modified.Newton"),
```

standardize.response = FALSE,
    type.multinomial = c("ungrouped", "grouped"))
{
    family = match.arg(family)
    alpha = as.double(alpha)
    this.call = match.call()
    nlam = as.integer(nlambda)
    y = drop(y)
    np = dim(x)
    nobs = as.integer(np[1])
    nvars = as.integer(np[2])
    taxonomy = as.data.frame(taxonomy)

    dimtax <- dim(taxonomy)
    if (dimtax[1] != nvars) stop(paste("number of variables does not match
        size of taxonomy"))
    if (!(dimtax[2] > 1)) stop(paste("taxonomy must consist of the otu
        level and at least one taxon level
        above"))

    if (family == "binomial" & is.null(dim(y))) {y = cbind(y, 1-y)}

    # Normalize covariates.
    if (standardize == TRUE) {x = scale(x)}
    else {x = scale(x, scale = rep(1,nvars))}

    modb = glmnet(x, y, family = family, offset = offset,
        alpha = alpha, lambda = lambda, standardize = FALSE,
        nlambda = nlam, thresh = thresh * 1e-01)
    ulam = modb$lambda
    nlam = length(ulam)

    # Initialize intermediates
    ntl = ncol(taxonomy)
    vnames = taxonomy[, ntl]
    taxa = lapply(taxonomy[, -ntl, drop = FALSE], levels)
    b_old = Matrix(rep(0,nvars), sparse = TRUE,
        dimnames = list(vnames, NULL))
    beta = list()
    b0 = c()

    for (i in 1:nlam) {
b_old = modb$beta[,i]

a_old = Matrix(rep(1,nvars), sparse = TRUE,
  dimnames = list(vnames, NULL))

d_old = lapply(taxa, function(x) data.frame(d=rep(1,length(x)),
  row.names=x))

red = redistribute6(taxonomy, d_old, a_old, b_old)

a_old = red$ao
d_old = red$do

w = phi(taxonomy, d_old) * abs(a_old)
w[w < 1e-2] = 0.5 * w[w < 1e-2] + 0.5 * 1e-2

for (j in 1:1000) {
  mod = glmnet(x, y, alpha = alpha, family = family,
    standardize = FALSE, lambda = ulam[i], penalty.factor = w^-1,
    thresh = thresh * 1e-01, maxit = maxit)
  b_new = mod$beta
  a_new = b_new * w^-1
  red = redistribute6(taxonomy, d_old, a_old, a_new)
  a_new = red$an
  a_old = red$ao
  d_new = red$do
  w = phi(taxonomy, d_new) * abs(a_new)
  expt = min(50, j + 2)
  w[w < 10^-expt] = 0.5 * w[w < 10^-expt] + 0.5 * 10^-expt

  diff1 = sum(abs(b_new - b_old))
  diff2 = sum(abs(b_new - b_old) / max(abs(b_old), thresh))
  b_old = b_new; d_old = d_new;

  print(paste("Lambda ", i, ", " = ", ulam[i], ", Iter ", j, ": ",
    diff1, diff2, sep=" "))
  if (diff2 < thresh) break
}

beta = c(beta, b_old)
b0 = c(b0, mod$a0)

beta = do.call(cBind, beta)
beta = beta / attr(x, "scaled:scale")
b0 = b0 - as.numeric(attr(x, "scaled:center")) *%*% beta

return(list(beta=beta, a0 = b0, lambda = ulam))
phi <- function(taxonomy, d = NULL, a = NULL)
  if (is.null(d)) return(a)
  D <- vapply(1:length(d), function(i) d[[i]][taxonomy[,i]], numeric(nrow(taxonomy))
  if (is.null(a)) return(apply(D,1,prod))
  b = apply(D,1,prod) * a; row.names(b) = row.names(a)
  return(b)

redistribute6 = function(taxonomy, do, ao, an) {
  taxn = do.call(c,lapply(do, rownames))
  taxonomy = do.call(cbind, lapply(taxonomy, as.integer))
  ntl = ncol(taxonomy)
  ni = nrow(taxonomy)
  ntax = vapply(do, nrow, integer(1))
  itax = rbind(as.integer(1),ntax)
  nta = as.integer(ntax[length(ntax)])
  itax = vapply(1:(ntl-1), function(i) {
    if (i==1) {itax[,i]}
    else {itax[,i] + ntax[i-1]}
  }, numeric(2))
  taxonomy = vapply(1:(ntl), function(i) {
    if (i==1 | i == ntl) {taxonomy[,i]}
    else {taxonomy[,i] + ntax[i-1]}
  }, numeric(ni))
  do = as.matrix(do.call(rbind,do)); row.names(do) = taxn
  
  red = .Fortran("redistribute6", taxo = as.integer(taxonomy),
              itax = as.integer(itax), dol = as.numeric(do),
              aol = as.numeric(ao), ane = as.numeric(an), ntl = ntl,
              ni = ni, nta = nta)
  an[] = red$ane
  ao[] = red$aol
  do[] = red$dol
  do = lapply(1:(ntl-1), function(i) do[itax[1,i]:itax[2,i],,drop=FALSE])
  
  return(list(do=do,ao=ao,an=an))
}
A.0.2 fortran-code

c File: redistribute6.f

c Version: 1.0

c Date: 2016-06-17

c Author: Stephen Rush <srush01@uoguelph.ca>

c Maintainer: Stephen Rush <srush01@uoguelph.ca>

c Description: Subroutine to redistribute mass among the group and individual coefficients according to mass equilibrium result

subroutine redistribute6(taxo, itax, dol, aol, ane, ntl, ni, nta)
    implicit none
    integer, parameter :: dp = kind(1.d0)
    integer :: ntl, ni, nta, i, j, k, jerr
    integer :: taxo(ni,ntl), itax(2,ntl-1)
    real :: att, dtt
    real(kind=dp) :: dol(nta), aol(ni), ane(ni)
    integer, dimension(:), allocatable :: cind

    allocate(cind(1:ni), stat = jerr)

    aol = sqrt(abs(aol) * abs(ane))

c Determine new alpha
    do i = (ntl-1),1,-1
        do j = itax(1,i),itax(2,i)
            att = 0
            do k = 1,ni
                if (taxo(k,i) /= j) cycle
                att = att + aol(k)
            enddo
            att = max(att ** (1.0/(ntl-i+2)), 1e-20)
            dtt = dol(j) ** (1.0/(ntl-i+2))
            do k =1,ni
                if (taxo(k,i) /= j) cycle
                aol(k) = aol(k) * dtt / att
            enddo
        enddo
    enddo

    ane = sign(aol, ane)

c Build new group coefficients
do i = 1,(ntl-1)
    do j = itax(1,i),itax(2,i)
        dtt = 0
        do k = 1,ni
            if (taxo(k,i) /= j) cycle
            dtt = dtt + aol(k)
        enddo
        dol(j) = dtt
    enddo
enddo

dereallocate(cind)
return
end