STATISTICAL DISCOVERY OF BIOMARKERS IN METAGENOMICS

by

Ahmad Hakeem Abdul Wahab

A Thesis Submitted to the Faculty of the

STATISTICS GRADUATE INTERDISCPLINARY PROGRAM

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2015

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Ahmad Hakeem Abdul Wahab

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Lingling An
Professor of Biometry

(8 July 2015)
Date

ACKNOWLEDGEMENT

I would like to extend a hearty gratitude to my advisor Dr. Lingling An who was willing to take me in and mentor me. Her unwavering support, encouragement and unyileding intellectual curiosity has made me grow and mature as both a researcher and a person. I would also like to thank my committee members Dr. Ning Hao and Dr. Bonnie Hurwitz who were invaluable in providing me with guidance and direction.

My thanks to previous lab members Ruofie Du and Naruekamol Pookhao and not forgetting the current members Meng Lu, Dan Luo, Brooke Rabe and especially Sara Ziebell who helped me with my presentations and provided insightful opinions regarding my research.

I am grateful for the support from my siblings who have been my pillars of strength in the many days where I was encumbered by stress. I have been so fortunate to have been blessed with amazing parents who have always been there for me, through thick and thin. No amount of thanks can describe how much I am thankful for their love and support.

CONTENTS

1 ABSTRACT

Metagenomics holds unyielding potential in uncovering relationships within microbial communities that have yet to be discovered, particularly because the field circumvents the need to isolate and culture microbes from their natural environmental settings. A common research objective is to detect biomarkers, microbes are associated with changes in a status. For instance, determining such microbes across conditions such as healthy and diseased groups for instance allows researchers to identify pathogens and probiotics. This is often achieved via analysis of differential abundance of microbes. The problem is that differential abundance analysis looks at each microbe individually without considering the possible associations the microbes may have with each other. This is not favorable, since microbes rarely act individually but within intricate communities involving other microbes.

An alternative would be variable selection techniques such as Lasso or Elastic Net which considers all the microbes simultaneously and conducts selection. However, Lasso often selects only a representative feature of a correlated cluster of features and the Elastic Net may incorrectly select unimportant features too frequently and erratically due to high levels of sparsity and variation in the data.

In this research paper, the proposed method AdaLassop is an augmented variable selection technique that overcomes the misgivings of Lasso and Elastic Net. It provides researchers with a holistic model that takes into account the effects of selected biomarkers in presence of other important biomarkers. For AdaLassop, variable selection on sparse ultra-high dimensional data is implemented using the Adaptive Lasso with p-values extracted from Zero Inflated Negative Binomial Regressions as augmented weights.

Comprehensive simulations involving varying correlation structures indicate that AdaLassop has optimal performance in the presence multicollinearity. This is especially apparent as sample size grows. Application of Adalassop on a Metagenome-wide study of diabetic patients reveals both pathogens and probiotics that have been researched in the medical field.

2 List of Figures

3 List of Tables

4 BACKGROUND

4.1 METAGENOMICS

Microbes are pervasive single-celled organisms that are crucial for everyday elementary bio-processes on earth. They convert key life-sustaining elements such as oxygen, carbon and nitrogen into forms biologically accessible to other higher order organisms. They reside on epithelials and within the human gut to protect us against other harmful microbes. They even help purify toxins in the environment that are both man-made and occur naturally. Life as we know it is sustained via the tumultuous undertaking of these microbes. These bio-processes and activities however are executed not by a single microbe, but by communities of microbes in their own natural environments. As such, there is growing interest to study microbial communities and how they occur naturally in their own respective environments.

Although genetic sequencing technologies have grown by leaps and bounds over recent decades, only 1% of microbial communities can be successfully cultured in controlled laboratory settings (Hugenholtz, 2002; Wooley and Ye, 2010). Enter Metagenomics, the study of microbial communities sampled directly from a given environment (eg. soil, human gut, ocean beds) circumventing the necessity for cloning and culturing each particular microbe to better understand the genomic diversity of microbes without perturbing the inherent homeostasis of their ecosystem (Hugenholtz, 2002; Huson *et al.*, 2009; Kunin *et al.*, 2008; Wooley and Ye, 2010).

One of the primary objectives of Metagenomic analysis is to make quality inferences about how these microbial communities differ. There are two main approaches to ascertain these differences: functional and taxonomic analysis. As their names imply, functional analysis looks at biological processes that occur in a system and compare their occurrences under different phenotypic (e.g. diseased, healthy or disparate treatments) or physical conditions (e.g. different locations in the human gut). Taxonomic analysis, on the other hand, achieves the same aim by quantifying the different abundances of genera or species of microbes from

the sampled environment. In this paper, taxonomic analysis is conducted to ascertain which bacterial strains are differentially abundant or expressed, i.e. exhibit a decrease or increase across different phenotypic conditions. These bacteria or features of interest are defined as *Differentially Abundant Features* or DAFs for short.

4.2 Background: Linear Models and Variable Selection

Due to the ultra high dimensionality of Metagenomic data, an efficient variable selection technique is necessary to comb through data for knowledge discovery as well as enhance the prediction performance of the selected model (Fan & Li, 2006). To better understand the process of Variable Selection, consider the classical linear model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \\ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix} \end{aligned}$$

where $\mathbf{x}_i'$ is the ith row (sample or individual i) and $\mathbf{X}_i$ is the ith column (bacterial strain or feature i) of the design matrix $\mathbf{X}$.

We assume the response $\mathbf{y}$ has been centered and the design matrix $\mathbf{X}$ has been standardized, thus excluding the intercept parameter for ease of interpretation, i.e.

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for } j = 1, \dots p$$

Furthermore, the errors are assumed to be independent and identically distributed with mean 0 and homogeneous finite variance σ^2. The elementary interest is in the sparse modeling problem where the true model has sparse coefficients, i.e. some components of $\boldsymbol{\beta}^*$ are exactly zero. The crux of variable selection is the discovery of the set of non-zero β-coefficients under

the sparse representation $\mathscr{A} = \{j : \beta_j^* \neq 0, j = 1, \ldots, p\}$. Simply put, variable selection informs researchers which x's or features are important, based on the given data.

4.2.1 LASSO

One of the earlier methods of variable selection using penalizing functions is the Lasso (Tibshirani, 1996). The Lasso estimator is calculated by optimizing the full regression model subject to the L_1 constraint:

$$\beta(\text{Lasso}) \quad = \quad \underset{\beta}{\arg\min} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}$$

Relative to traditional unbounded least square estimation, the L_1 penalty allows the model to regularize the least square regression while inducing sparsity, shrinking the number of features in the fitted model.

However, the Lasso has two major deficiencies: Lasso solutions lack the oracle property and yield unstable solutions in high dimensional data. A selected model is said to have the oracle property (Fan & Li, 2001) if the β-coefficients attain asymptotic normality and exhibit consistency in variable selection, i.e. the probability that the model correctly selects important features tends to 1.

4.2.2 ADAPTIVE LASSO

To overcome the deficiencies of Lasso, Zou (2006) proposed the Adaptive Lasso estimator, which was shown to perform comparably to the oracle. The Adaptive Lasso is defined as:

$$\beta(\text{AdaLasso}) \quad = \quad \underset{\beta}{\arg\min} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| \right\}$$

The weights $\hat{\omega}_j$'s are computed adaptively via the relation $\hat{\omega}_j = |\hat{\beta}_j^{\text{Ini}}|^{-\gamma}$, where γ is some positive tuning parameter and $\hat{\beta}^{\text{Ini}}$ is the initial root-n estimate consistent of true β's.

4.2.3 Elastic Net

As mentioned before, Lasso performs sub par in high dimensional data due to the presence of multicollinearity, i.e. when the input x-variables are correlated. For example, there are 3 differentially abundant features that are highly correlated in the fitted regression model. Lasso tends to select one variable as a representative of the correlated cluster, ignoring the remaining two or forcing their β coefficients to be 0.

Zou and Hastie (2005) proposed the Elastic Net as a boosted Lasso that performs well in high dimensional data by omtroducing a second penalty term λ_2. The Elastic Net estimator is defined as:

$$\beta(\text{Enet}) \quad = \quad \left(1+\frac{\lambda_2}{n}\right)\underset{\beta}{\arg\min}\left\{||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2+\lambda_2||\beta||_2^2+\lambda_1||\beta||_1\right\}$$

In summary, the L_1-penalty is responsible for variable selection by inducing sparsity while the L_2-penalty stabilizes the solution path, thus improving prediction performance.

4.2.4 Adaptive Elastic net

Although the Elastic Net performs well in high dimensions, it does not have the oracle property. Recall that the Adaptive Lasso performs comparably to a model that has oracle property. A natural proposal would thus be to combine the Elastic net and the Adaptive Lasso. Zou and Zhang (2009) proposed the Adaptive Elastic Net which has both the oracle property and handles multicollinearity well. The Adaptive Elastic Net Estimator is defined as:

$$\beta(\text{AdaEnet}) \quad = \quad \left(1+\frac{\lambda_2}{n}\right)\underset{\beta}{\arg\min}\left\{||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2+\lambda_2||\beta||_2^2+\lambda_1^*\sum_{j=1}^{p}\hat{\omega}_j|\beta_j|\right\}$$

The weights $\hat{\omega}_j$'s are computed adaptively via the relation $\hat{\omega}_j = |\hat{\beta}(\text{Enet})_j|^{-\gamma}$, where γ is some positive tuning parameter, $\hat{\beta}(\text{Enet})_j$ are the Elastic Net estimators and λ_1 and λ_2 are the L_1 and L_2 penalties respectively. The γ's and weights $\hat{\omega}_j$'s are chosen such that larger weights are

assigned to features that are not important and smaller weights are assigned to DAFs.

4.3 MOTIVATION

Typically, one would employ a screening method prior to using these adaptive models such as the Sure Independent Screening (SIS) (Fan and Lv, 2008). Implementing these adaptive models in situations where the number of features is much larger than the number of samples ($p >> n$) without any prior screening mechanism would yield an adaptive model that is just as good as the regular Lasso or Elastic Net.

In this paper, the screening step is skipped the whole procedure of variable selection and biomarker discovery is implemented in two steps: the adaptive weights estimation and the variable selection step. In the adaptive weights estimation, a Zero Inflated Negative Binomial Regression is conducted on each feature against the dichotomous conditions to calculate the p-value for each feature. In the variable selection step, these p-values are then used in place of the adaptive weights. Note that this step is also the variable selection step since only selected variables have non-zero coefficients while others are forced to zeros. The coefficients are then calculated and fitted into a Logistic Regression predictive model for the purposes of Statistical Classification.

It is worth emphasizing that the AdaLassop does not transform data into a normal distribution or use proportions and simply works on raw data. AdaLassop simply integrates the use of p-values derived from the initial adaptive weights estimation into the variable selection model without imposing any prespecified significance level α. Relative to the original Adaptive Lasso and Adaptive Elastic Net, AdaLassop does not require estimating the γ parameter in the calculation of the adaptive weights. Finally, Adaptive Lasso is chosen over the Adaptive Elastic Net since simulations and real data analysis indicate that the selection performance of both Elastic Net and Adaptive Elastic Net with Metagenomics data are unstable.

The outline of this paper is as follows: Section 2 explains the pipeline of the proposed method AdaLassop, Section 3 describes the simulation studies and results, Section 4 demon-

strates an application of the method to a diabetes dataset and Section 5 closes with a discussion on the proposed method.

5 METHODOLOGY

Three defining features of Metagenomics abundance data are: high frequencies of zero, overdispersion (variance of data far outweighs the mean) and ultra high dimensionality in presence of high levels of sparsity ($p >> n$) (Pookhao *et al.*, 2014). For instance, observations of feature counts from the Metagenome-wide assoication study from Qin,J. *et al.* (2012) have the following structure:

FeatureID	43	244	490
Healthy	0,0,0,0,0,0,0, 1101,1030,0,0,0, 491,1181,0,0,0,0, 0,0,0,0,0,0,0	0,0,0,0,228429, 0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0, 0,7154,0,0,0	1191,0,3235,2379,3363,3099, 2208,3655,2478,2567,2965,1013, 2487,4209,2006,1681,3326,3481, 1213,0,0,4638,1953,1887,0
Diabetic	0,0,0,0,0,0,0,0, 649,0,0,0,0, 0,0,0,0,0,0,0,0, 0,0,402,762	0,0,0,0,0,0,0,0, 1278,0,0,1742, 0,107485,0,0,0, 0,0,0,0,0,0,0,0	2006,1227,0,1296,2350,2552,0, 0,3855,1699,1193,0,836,1946, 1426,2759,0,1866,34891,8960, 1841,3605,1243,2207,2476

Table 5.1: Observation examples from Diabetes Metagenomics Dataset (Qin,J. *et al.*, 2012) with 25 samples per group

A Zero Inflated Negative Binomial distribution (Rapaport *et al.*, 2013) or a Zero Inflated Generalized Poisson (Tuenter, 2000) for instance would model Metagenomic data more effectively than the more popularized Poisson for conventional count data due to their ability to model overdispersion. Thus, proposed statistical methods in the realm of Metagenomics should be able to factor in zero-inflation, overdispersion and multicollinearity to allow for quality inferences.

5.1 STEP 1: ADAPTIVE WEIGHTS ESTIMATION

5.1.1 NORMALIZATION

One of the main concerns in high thoroughput sequencing technologies is the arbitrary number of reads that are generated with high degrees of variation across samples that arise from the sampling process (Pookhao N, *et al.*, 2015). This is primarily attributed to the

disparate sequencing depths of the read counts across multiple samples. For this very reason, metagenomics data need to be pre-processed to account for this source of bias or variation. This step is more commonly known as normalization (not to be confused with transforming a data to have it follow a normal distribution), which scales the data across samples to make them more comparable. The Trimmed Method of M-values is employed (Robinson and Oshlack, 2010) to normalize the data, which is implemented in the edgeR Bioconductor package of R.

5.1.2 Zero Inflated Negative Binomial Regression

The p-values are then extracted from the Zero Inflated Negative Binomial Regression model to adaptively calculate the weights in the AdaEnetp model. Granted, there are many alternatives to model count data with overdispersion (Hausman, Hall and Griliches, 1984), Cameron and Trivedi, 1998). Readers are directed to (Green, 2003) and (Winkelmann, 2008) for further reading on regression models for count data.

As a start, consider the Poisson Regression Model, since the Negative Binomial Regression is simply a generalization of the former. Assume the following linear model, with responses y_i and regressors $\mathbf{x}_i$, where $\mathbf{x}_i$ is the i-th row of the design matrix $\mathbf{X}$. Then the primary equation of out model or Probability Mass Function (PMF) is denoted by

$$P(Y = y_i|\mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\Gamma(y_i+1)},\ y_i = 0,1,2,\ldots$$

The most common link function for λ_i is denoted by the loglinear model.

$$\log\lambda_i = \mathbf{x}_i'\boldsymbol{\beta}$$

This specification works well with the Poisson model, since the expectation is the same as the variance,

$$E[y_i|\mathbf{x}_i] = Var[y_i|\mathbf{x}_i] = \lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$$

Count data often exhibit overdispersion, i.e. when variance far outweighs the expectation. To rectify this, the Poisson model is generalized by introducing an error or unobserved effect ϵ_i to the unconditional mean

$$\begin{aligned} \log\gamma_i &= \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i = \log\lambda_i + \log u_i \\ \gamma_i &= \lambda_i u_i \end{aligned}$$

By this specification, the distribution of $(y_i|\mathbf{x}_i, u_i)$ still follows a Poisson distribution Poi(γ_i):

$$P(Y = y_i|\mathbf{x}_i, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{\Gamma(y_i+1)},\ y_i = 0, 1, 2, \ldots$$

Since the above expression is conditional on the specification errors u_i, then the conditional distribution of y_i given only $\mathbf{x}_i$ is found by integrating $P(Y = y_i|\mathbf{x}_i, u_i)$ over u_i:

$$\begin{aligned} f(y_i|\mathbf{X}_i) &= P(Y = y_i|\mathbf{x}_i) \\ &= \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{\Gamma(y_i+1)} g(u_i)\,du_i \end{aligned}$$

For mathematical convenience, assume $u_i = e^{\epsilon_i} \sim$ Gamma(θ, θ). Then the Probability Mass Function (PMF) is given by:

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{u_i\theta}$$

The density of y_i with the accompanying expectation and variance is given as:

$$
\begin{aligned}
f(y_i|\mathbf{x}_i) &= P(Y = y_i|\mathbf{x}_i) \\
&= \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{\Gamma(y_i+1)} g(u_i)\,du_i \\
&= \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{\Gamma(y_i+1)} \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{u_i\theta}\,du_i \\
&= \frac{\theta^\theta}{\Gamma(\theta)} \frac{{\lambda_i}^{y_i}}{\Gamma(y_i+1)} \int_0^\infty u_i^{(y_i+\theta)-1} e^{-(\lambda_i+\theta)u_i}\,du_i \\
&= \frac{\theta^\theta}{\Gamma(\theta)} \frac{{\lambda_i}^{y_i}}{\Gamma(y_i+1)} \Gamma\left(y_i+\theta\right)(\lambda_i+\theta)^{-(y_i+\theta)} \\
&= \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)} \left(\frac{\lambda_i}{\lambda_i+\theta}\right)^{y_i} \left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta} \\
&= \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)} r_i^{y_i}(1-r_i)^{\theta}, \text{ where } r_i = \frac{\lambda_i}{\lambda_i+\theta} \\
E(y_i|\mathbf{x}_i) &= \lambda_i \\
Var(y_i|\mathbf{x}_i) &= \lambda_i\left(1+\frac{\lambda_i}{\theta}\right) \\
\lambda_i &= e^{\mathbf{x}_i'\boldsymbol{\beta}}
\end{aligned}
$$

The expression above defines a Negative Binomial Type 2 (Cameron and Trivedi, 1998) distribution. The Negative Binomial Type 1 is attained if $\theta_i = \theta\lambda_i$, reducing r_i to $r = \frac{1}{1+\theta}$. The density of the Negative Binomial Type 1 is thus

$$
\begin{aligned}
f(y_i|\mathbf{x}_i) &= \frac{\Gamma(\theta\lambda_i+y_i)}{\Gamma(y_i+1)\Gamma(\theta\lambda_i)} r^{y_i}(1-r)^{\theta\lambda_i}, \text{ where } r = \frac{1}{1+\theta} \\
E(y_i|\mathbf{x}_i) &= \lambda_i \\
Var(y_i|\mathbf{x}_i) &= \lambda_i\left(1+\frac{1}{\theta}\right)
\end{aligned}
$$

There is no theoretical evidence to support the preference of one model over the other, other than most statistical packages use Type 2 for implementation (Green, 2008). A notable difference between the two models is that Type 2 has a variance that is quadratic in the mean while

Type 1 is not. We use the NB-2 regression for this study.

Since Metagenomic data is directly sampled from the environment, one cannot rule out the possibility of the zeros arising from sampling error. A more appropriate assumption would be that the zero counts from the data come from "structural" and "sampling" origins. Hence, a zero inflated model is deemed more appropriate than a Hurdle model (Mullahy, 1986; Hu *et al.*, 2011) which assumes that all zeros come from a structural source. A Zero Inflated Negative Binomial Regression can be obtained by mixing a probability distribution degenerate τ_i at 0 with the Negative Binomial Regression, yielding the following PMF with the accompanying expectation and variance (Ismail and Zamani, 2013):

$$f(y_i|\mathbf{x}_i) = \begin{cases} \tau_i + (1-\tau_i)(1-r_i)^{\theta} & y_i = 0 \\ (1-\tau_i)\frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)} r_i^{y_i}(1-r_i)^{\theta} & y_i \neq 0 \end{cases}$$

$$E(y_i|\mathbf{x}_i) = (1-\tau_i)\lambda_i$$

$$Var(y_i|\mathbf{x}_i) = (1-\tau_i)\lambda_i\left(1+\frac{\lambda_i}{\theta}+\tau_i\lambda_i\right)$$

$$r_i = \frac{\lambda_i}{\lambda_i+\theta}$$

The Zero Inflated Negative Binomial Regression above can be estimated using Maximum Likelihood.

In this case, the response would be the individual features $\mathbf{X}_i$ (columns of the design matrix $\mathbf{X}$) and the input would be the binary variable $\mathbf{y}$ for the disease status or condition indicator. In the context of a Generalized Linear Model (GLM), a convenient way to link the response variables to the input variables would be to use a loglink relationship. The regression equation for the mean of feature i with respect to the input binary condition $\mathbf{y}$ from the main model (Achim *et al.*, 2008) is thus given by

$$\mu_i = \tau_i * 0 + (1-\tau_i)e^{\mathbf{y}'\boldsymbol{\beta}} \tag{5.1}$$

5.1.3 Multiple Testing

Hypotheses tests are conducted for each feature against the disease status from the Zero-Inflated Negative Binomial Regression models (equation 5.1):

$$H_{0,j} : \beta_j = 0 \quad \text{vs} \quad H_{1,j} : \beta_j \neq 0,\ j = 1, \dots p$$

The test statistic z follows a standard normal distribution and is used to test against a two-sided alternative that the estimated coefficient $\beta_j \neq 0$. The corresponding p-value is the probability of observing this z test statistic value assuming that the null is true. Thus, a small p-value is favorable since it is indicative that there is a small probability of observing this data, assuming there is no real difference of that the null is true.

These raw p-values are then adjusted for simultaneous hypothesis testing using Benjamini-Hochberg's procedure (Benjamini and Hochberg, 1995). The procedure is as follows:

1. Sort and rank the p-values, where 1 is the rank of the smallest p-value.
2. Multiply each p-value by the total number of p-values (m) divided by the rank of the assigned p-value (k).
3. Take the minimum of the adjusted p-value and 1
4. Finally, take the minimum of the resultant adjusted p-value and the adjusted p-values of the remaining higher raw p-values.

These adjusted p-values are thus defined as

$$\tilde{p}_j = \min_{k \in \{i, \dots m\}} \left\{ \min \left\{ \frac{m}{k} p_{(k)}, 1 \right\} \right\}$$

5.2 Step 2: Variable Selection

5.2.1 Logistic Regression

The logistics regression is adopted for a binary response for the model (Johnston, 1971). Given the responses are binary variables: -1 for healthy and 1 for diabetic individuals,

$$\begin{aligned}
\log\left(\frac{P(y=1|\mathbf{x})}{1-P(y=-1|\mathbf{x})}\right) &= \mathbf{x}'\boldsymbol{\beta} \\
P(y=1|\mathbf{x}) &= \frac{e^{-\mathbf{x}'\boldsymbol{\beta}}}{1+e^{-\mathbf{x}'\boldsymbol{\beta}}} = \frac{1}{1+e^{-\mathbf{x}'\boldsymbol{\beta}}} \\
P(y=-1|\mathbf{x}) &= 1-P(y=1|\mathbf{x}) = \frac{1}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}
\end{aligned}$$

The interpretation of the coefficients follows directly: a positive coefficient means that as the feature count increases, it increases the probability of an individual having diabetes. Conversely, a negative coefficient would mean that as the feature count increases, it decreases the probability of an individual not having diabetes.

Since there are only two conditions, the decision boundary is essentially the hyperplane that satisfies the equation $\mathbf{x}\boldsymbol{\beta} = 0$. The predicted status is thus given by

$$\hat{y}_i = \begin{cases} 1 & , \frac{1}{1+e^{-\mathbf{x}'\boldsymbol{\beta}}} \geq 0 \\ -1 & , \frac{1}{1+e^{-\mathbf{x}'\boldsymbol{\beta}}} < 0 \end{cases}$$

Assuming y is the label for the feature vector $\mathbf{x}$, the accompanying likelihood and log-likelihood functions are thus given by

$$\begin{aligned}
L(y,\mathbf{x},\boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{1}{1+e^{-y_i\mathbf{x}_i'\boldsymbol{\beta}}} \\
l(y,\mathbf{x},\boldsymbol{\beta}) &= \arg\min_{\beta}\left\{\sum_{i=1}^{n}\log\left(1+e^{-y_i\mathbf{x}_i'\boldsymbol{\beta}}\right)\right\}
\end{aligned}$$

This formally defines the logistic loss function

5.2.2 AdaLassop & AdaLassobh

Using the logistic loss function from the previous section, the following penalized log-likelihood equation will be optimized:

$$\arg\min_{\beta} \left\{ l(y, \mathbf{x}, \boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} \hat{\omega}_i |\beta_j| \right\}$$

We note that above looks similar to the Adaptive Lasso model, with the only difference being the adaptive weights. AdaLassop is obtained when the adaptive weights $\hat{\omega}_i$ are the raw p-values while AdaLassobh is obtained when the adaptive weights $\hat{\omega}_i$ are the Benjamini-Hochberg adjusted p-values.

5.2.3 AdaEnetp & AdaEnetbh

As before, the following penalized log-likelihood equation will be optimized:

$$\arg\min_{\beta} \left\{ l(y, \mathbf{x}, \boldsymbol{\beta}) + \lambda_2 ||\boldsymbol{\beta}||_2^2 + \lambda_1 \sum_{j=1}^{p} \hat{\omega}_i |\beta_j| \right\}$$

We note that the above looks exactly like the Adaptive Elastic Net model, just that the adaptive weights are calculated differently. AdaEnetp is obtained when the adaptive weights $\hat{\omega}_i$ are the raw p-values while AdaEnetbh is obtained when the adaptive weights $\hat{\omega}_i$ are the Benjamini-Hochberg adjusted p-values.

6 Simulation Studies

6.1 Simulation Settings

1000 features comprising 25 and 100 samples for each condition each with 20 and 100 DAFs are generated. DAFs are defined as having different means across different conditions. The DAFs are generated from a a Zero Inflated Generalized Poisson (ZIGP) distribution. In the simulations, the mean for a DAF in one condition was set to be μ_i and $c_i\mu_i$ in the other condition, where μ_i is sampled from an interval $[1000, 2500]$. The dispersion parameter for each DAF is defined as the mean specified in condition 1 divided by 100.

The PMF of a Zero-Inflated Generalized Poisson (Consul and Pain 1970) is given

$$ZIGP(Y = y_i|\mu_i, \varphi_i, \tau_i) = \begin{cases} \tau_i + (1-\tau_i)e^{-\frac{\mu_i}{\varphi_i}} & y_i = 0 \\ (1-\tau_i)\left(\frac{1}{\Gamma(y_i+1)}\right)\mu_i(\mu_i + (\varphi_i - 1)y_i)^{y_i-1}\varphi^{-y_i}e^{-\frac{1}{\varphi_i}(\mu_i+(\varphi_i-1)y_i)} & y_i \neq 0 \end{cases}$$

with the corresponding expectation and variance

$$\begin{aligned} E(Y) &= (1-\tau_i)\mu_i \\ Var(Y) &= E(Y)(\varphi_i^2 + \mu_i\tau_i) \end{aligned}$$

The non-DAFs on the other hand are defined as having the same mean across different conditions. The non-DAFs are generated from a Zero-Inflated Negative Binomial. The non-DAF means μ_r are randomly selected from an interval $[1000, 5000]$ across both conditions. The dispersion for the ZINB (θ from ZINB model from section 2.2) is randomly selected from the interval $[0.1, 0.5]$. Note that non-DAFs that are correlated with the DAFs are generated using a ZIGP

Inflation of zeros is obtained via the τ_i parameter, which is the proportion of zeros in the simulate dataset. τ_i is randomly selected among 50%, 60% and 70%. Simulation settings are summarized in table 3.1.

Variable	Description	Range of Value
	DAF	20,100
	Samples/group	25,100
μ_i	DAF Mean	int[1000,2500]
μ_r	non-DAF Mean	int[1000,5000]
c_i	Mean effect	unif[2,4]
θ	Dispersion for ZINB	unif[0.1,0.5]
φ_i	Dispersion for ZIGP	$\sqrt{1+\frac{\mu_{\text{condition 1}}}{runif[0.1,10]}}$
τ_i	Prob of zero	0.5,0.6,0.7

Table 6.1: Description of simulation settings

In table 3.2, the left column represent information on condition 1 and the right column represents information on condition 2. The first row shows the overall means, while the second row shows the means over a smaller range and the third row shows the proportion of zeros in the data.

The rationale for using the stated simulation settings in table 3.1 stems from observations from the real data, shown in table 3.2. The means of the nonzero counts for features in condition 1 reach up to 250000 (row 1 column 1) while the means of the non-zero counts for features in condition 2 reach up to 300000 (row 1 column 2). However, there is a higher frequency of means in the range of 50000 counts and below for both conditions. Upon closer inspection of the distribution of means for both conditions up to 50000 count (row 2), the means have the highest frequencies in the range up to 5000 counts. Hence, the means of the non-DAFs are set to be in the range of $[1000,5000]$. The DAF means are set between $[1000,2500]$ since a mean effect factor or scaling factor c_i is to be selected from $[2,4]$, resulting in DAF means to fall in the $[2000,10000]$ range.

The motivation for the range of the dispersion parameter $\theta \in [0.1,0.5]$ for the ZINB stems from a literature overview of the supplementary file provided by (Sohn, *et al.*, 2015)

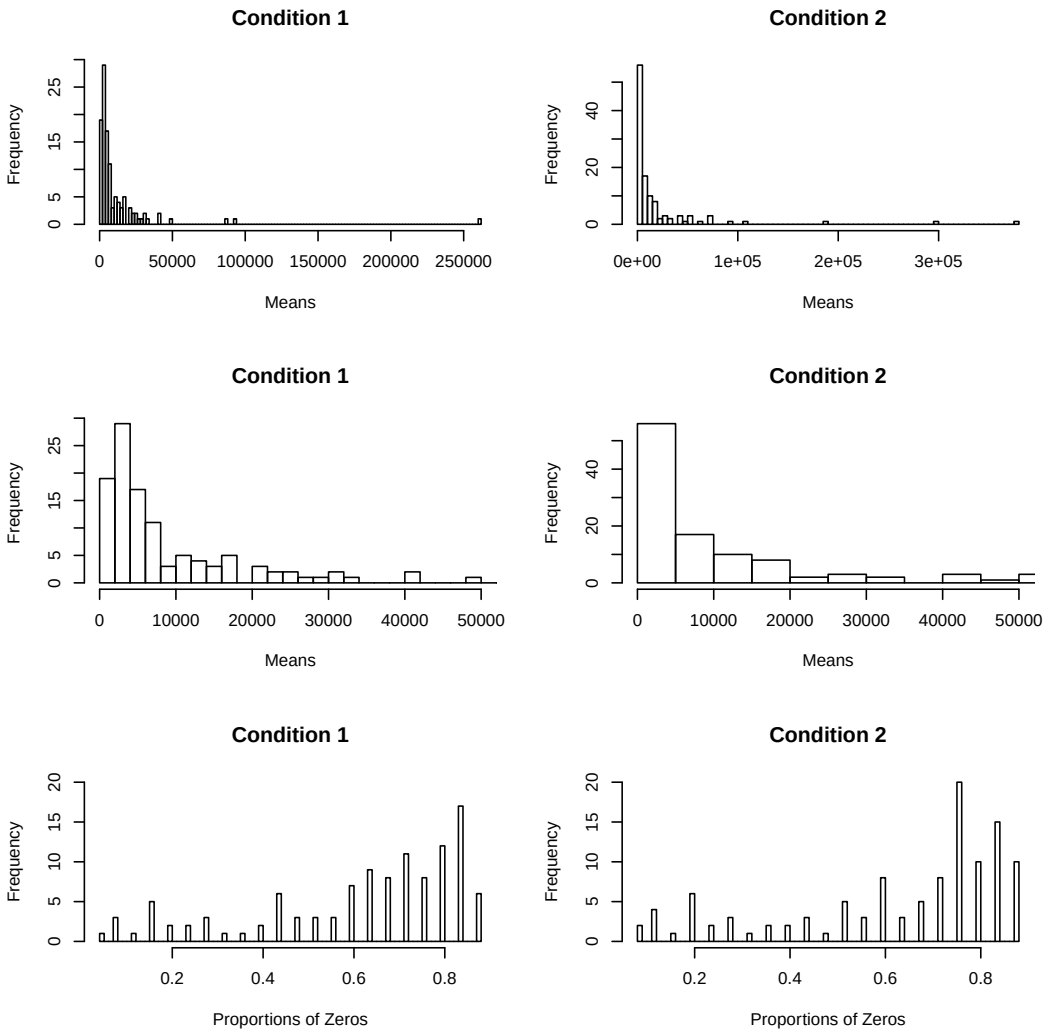

Table 6.2: Diagnostics of Real Data

6.1.1 DESIGN 1

There are 2 different clusters of features in Design 1: the DAF cluster (features 1-20) and the remaining non-DAF cluster (features 21-1000). Each DAF follows a Zero Inflated Generalized Poisson while the non-DAF follow a Zero Inflated Negative Binomial. No pre-existing correlation structure was specified between the DAFs and non-DAFs. The simulations are repeated for when there are 100 DAFs for sample sizes 25 and 100 respectively. The simulation settings

for Design 1 for 20 DAFs are summarized in table 3.3.

	DAF					Uncorr non-DAF		
Feature	1	2	...	19	20	21	...	1000
Condition1	μ_1	$c_1\mu_1$	...	μ_{10}	$c_{10}\mu_{10}$	μ_r	...	μ_r
Condition2	$c_1\mu_1$	μ_1	...	$c_{10}\mu_{10}$	μ_{10}			
Dispersion	$0.01\mu_1$	$0.01c_1\mu_1$	...	$0.01\mu_{10}$	$0.01c_{10}\mu_{10}$	θ		

Table 6.3: Settings for Design 1 for 20 DAFs

6.1.2 DESIGN 2

There are 3 different clusters of features in Design 2: the DAF cluster (features 1-20), the non-DAF cluster that is correlated to the DAF cluster (features 21-40) and the remaining uncorrelated non-DAF (features 41-1000). Unlike other studies, the DAFs simulated in Design 2 are actually correlated with the non-DAFs in an attempt to manually induce multicollinearity. In this case, feature 1 is correlated to feature 21, feature 2 is correlated to feature 22 until finally feature 20 is correlated to feature 40. Simulations are repeated for when there are 100 DAFs for sample sizes 25 and 100. In this case, the first cluster comprises features 1-100, the second cluster comprises features 101-200 and the final cluster comprises features 201-1000.

Correlation levels of 0.8 were induced for features in cluster 1 and cluster 2 across conditions in the following manner using copulas from the (corcounts) package in (Erhardt and Czado, 2009)R:

1. Generate a DAF with mean μ_i and a non-DAF with mean μ_r with a correlation of 0.8 for condition 1.

2. Generate a DAF with mean $c_i\mu_i$ and a non-DAF with the same mean μ_r from step (1) with a correlation of 0.8 for condition 2.

3. Bind the vectors of features from steps (1) and (2) to get a DAF that is correlated to a non-DAF at correlation level of around 0.8.

Simulations show that step (3) introduces randomness in the correlation structure. Empirically, the resultant correlation between cluster 1 and cluster 2 ranges between 0.65 and 0.8. The simulations are repeated for when there are 100 DAFs for sample sizes 25 and 100 respectively. The simulation settings for Design 2 for 20 DAFs are summarized in table 3.4.

	DAF					Corr non-DAF			non-DAF		
Feature	1	2	...	19	20	21	...	40	41	...	1000
Condition1	μ_1	$c_1\mu_1$	...	μ_{10}	$c_{10}\mu_{10}$	μ_r	...	μ_r	μ_r	...	μ_r
Condition2	$c_1\mu_1$	μ_1	...	$c_{10}\mu_{10}$	μ_{10}						
Dispersion	$0.01\mu_1$	$0.01c_1\mu_1$	...	$0.01\mu_{10}$	$0.01c_{10}\mu_{10}$	$0.01\mu_r$			θ		

Table 6.4: Settings for Design 2 for 20 DAFs

6.1.3 DESIGN 3

There are 3 different cluster of features in Design 3: the DAF cluster (features 1-10), the DAF cluster that is correlated to the first DAF cluster (features 11-20) and the remaining uncorrelated non-DAF (features 101-1000). In this case, feature 1 is correlated to feature 51, feature 2 is correlated to feature 52 until finally feature 50 is correlated to feature 100. Specifically for Design 3, simulations for 100 DAFs with 50 samples per group over 25 repetitions were conducted. The objective is to test if the AdaLassop inherits the Lasso's deficiency in selecting a representative feature from a correlated cluster.

Similar to Design 2, correlation was induced between DAF cluster 1 and DAF cluster 2 at within each condition, resulting in a correlation structure between features across conditions of between 0.65 to 0.8. The simulations are repeated for when there are 100 DAFs for sample sizes 25 and 100 respectively. The simulation settings for Design 3 for 20 DAFs are summarized in table 3.4.

6.2 SIMULATION RESULTS

Each simulation is repreated 25 times and the performance of each method is measured using True Positive Rates (TPR), False Positive Rates (FPR) and Misclassification Rates. In this

	DAF			Corr DAF			non-DAF		
Feature	1	...	10	11	...	20	21	...	1000
Condition1	μ_1	...	$c_5 mu_5$	μ_6	...	$c_{10}\mu_{10}$	μ_r	μ_r	μ_r
Condition2	$c_1\mu_1$	...	μ_5	$c_6\mu_6$	...	μ_{10}			
Dispersion	$0.01\mu_1$	...	$0.01c_5 mu_5$	$0.01mu_6$	...	$0.01c_{10} mu_{10}$	θ		

Table 6.5: Settings for Design 3 for 20 DAFs

study, TPR and FPR are defined as

$$\begin{aligned} \text{TPR} &= \text{Proportion of DAFs correctly selected} \\ \text{FPR} &= \text{Proportion of non-DAFs incorrectly selected} \end{aligned}$$

For instance, if there were 100 DAFs and only 80 were selected, then TPR would be 0.8. Subsequently, if there were 900 non-DAFs and 90 were selected, then FPR would be 0.1. The misclassification rate is simply the proportion of samples or individuals whose conditions are incorrectly estimated or predicted with the calculated coefficients. Thus, high TPRs with low FPRs and Misclassification Rates are ideal.

The objective function in the R package (gcdnet) is defined as

$$\frac{1}{N} Loss(y, X, \beta) + \frac{1}{2}\lambda_2 ||\beta||_2^2 + \lambda_1 \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|$$

For computational convenience, λ_2 was set to be either 1 or 0 and only λ_1 was estimated using a five-fold cross validation. Table 3.6 lists the 6 methods of comparisons for our simulation studies with their corresponding model settings:

	Adaptive Weights $\hat{\omega}_i$	λ_2
Lasso	1	0
AdaLassop	Raw p values	0
AdaLassobh	BH-adjusted p values	0
Enet	1	1
AdaEnetp	Raw p values	1
AdaEnetbh	BH-adjusted p values	1

Table 6.6: Methods for comparisons

6.2.1 DESIGN 1

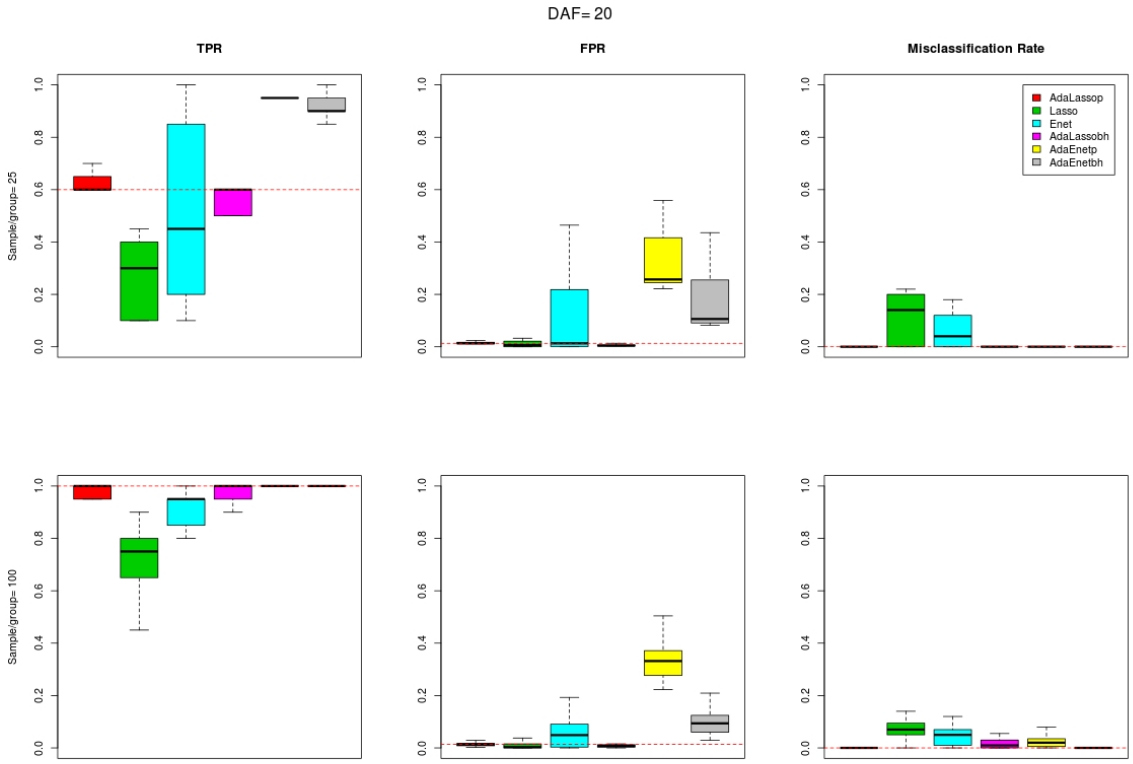

Figure 6.1: Boxplots for Design 1 with 20 DAFs

Figure 3.1 shows True Positive Rates in the first column, False Positive Rates in the second column and Misclassification Rates in the third column for sample sizes 25 per group on the top row and 100 per group on the bottom row over 25 repetitions for 20 DAFs. The red dotted

line indicates the median performance for the AdaLassop for all measures.

When N=25 per group, the TPR for AdaLassop has average performance and is outperformed by AdaEnetp and AdaEnetbh by a margin of almost 30%. However, AdaLassop controls FPR at an almost zero rate, unlike AdaEnetp and AdaEnetpbh which have 10% and 25% FPR respectively. Only Lasso and Enet have nonzero Misclassification Rates. When N=100 per group, the AdaLassop outperforms all other methods in terms of TPR, FPR and Misclassification Rate, while AdaEnetp and Adaenetpbh sustain similar median FPRs.

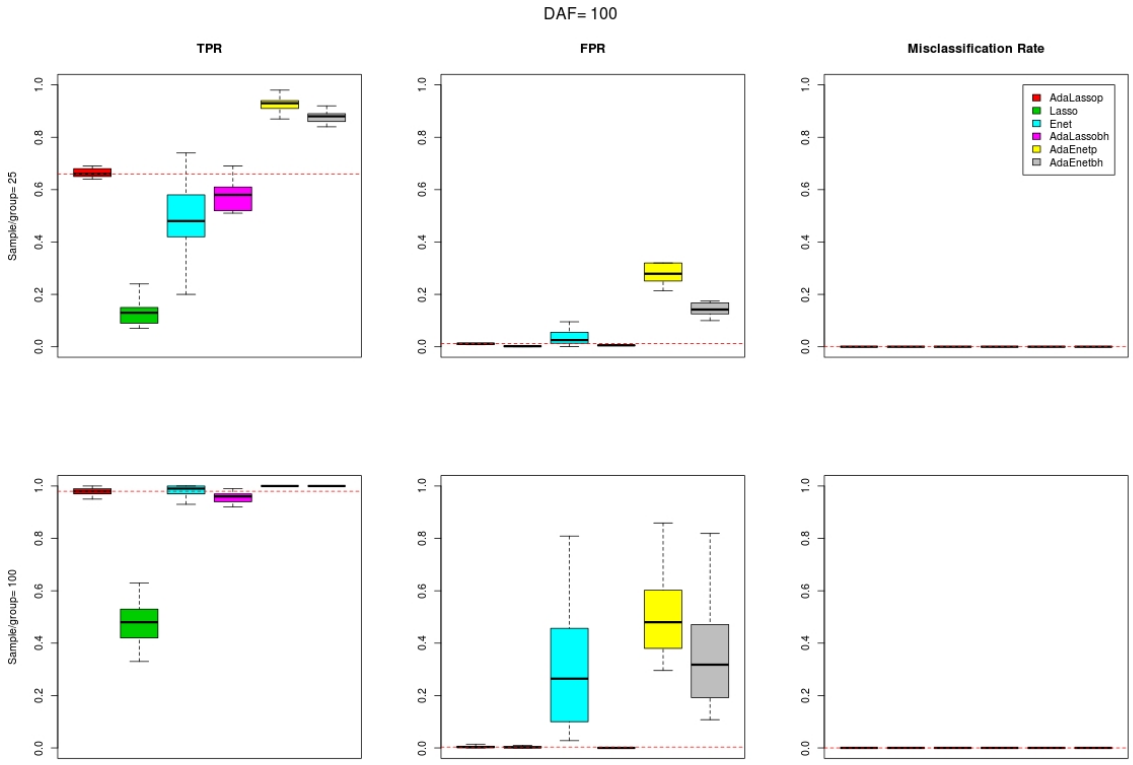

Figure 6.2: Boxplots for Design 1 with 100 DAFs

Figure 3.2 is set up as figure 3.1 but for 100 DAFs. When N=25 per group, the TPR for AdaLassop has average performance and is outperformed by AdaEnetp and AdaEnetpbh by a margin of almost 20%. However, AdaLassop controls FPR at an almost zero rate, unlike AdaEnetp and AdaEnetbh which have 25% and105% FPR respectively. All methods have zero

Misclassification Rates. When N=100 per group, the AdaLassop has an almost perfect TPR while maintaining a zero FPR and Misclassification Rate, while Enet, AdaEnetp and Adaenetbh suffer huge variations in FPR.

6.2.2 DESIGN 2

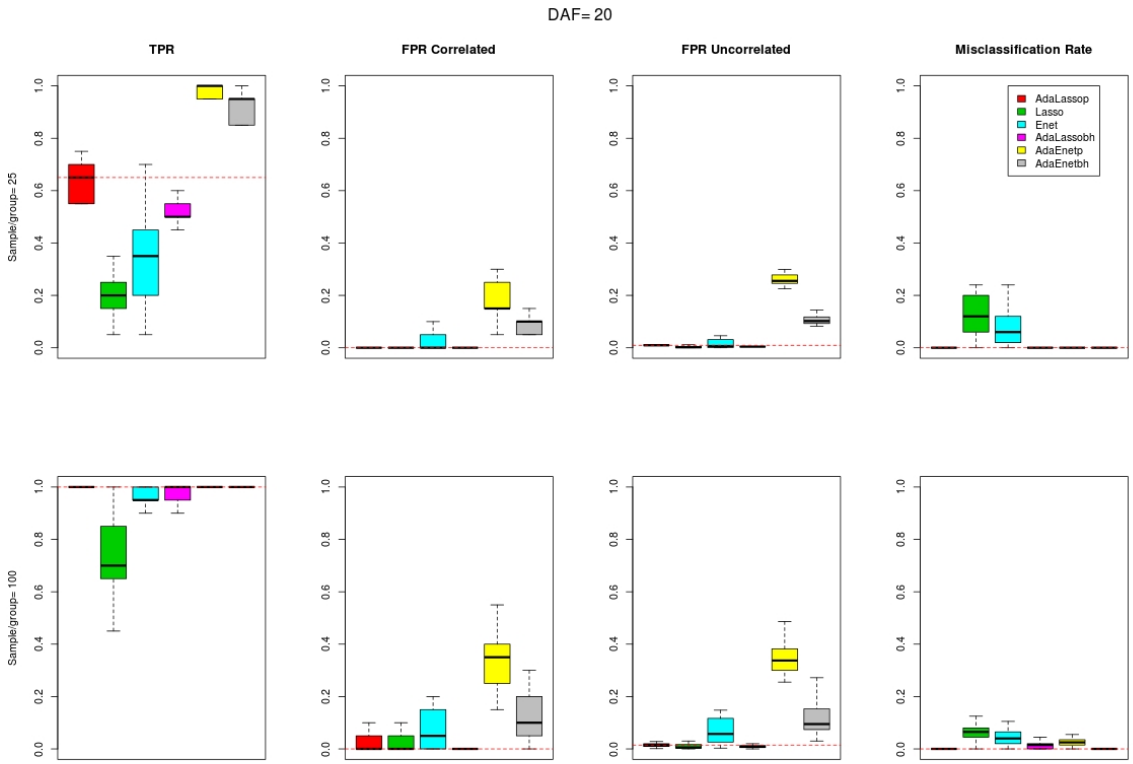

Figure 6.3: Boxplots for Design 2 with 20 DAFs

Figure 3.3 shows True Positive Rates in the first column, False Positive Rates for correlated nonDAFs in the second column, False Positive Rates for uncorrelated nonDAFs in the third column and Misclassification Rates in the last column for sample sizes 25 per group on the top row and 100 per group on the bottom row over 25 repetitions for 20 DAFs. The red dotted line indicates the median performance for the AdaLassop for all measures.

In design 2, DAF 1 is correlated to non-DAF 21, DAF 2 is correlated to non-DAF 22 and so on.

When N=25 per group, the TPR for AdaLassop has average performance and is outperformed by AdaEnetp and AdaEnetpbh by a margin of almost 30%. However, AdaLassop controls FPR for the correlated non-DAFs at an almost zero rate, unlike AdaEnetp and AdaEnetpbh which have 20% and 10% FPR respectively. Further, AdaLassop controls FPR for the uncorrelated non-DAFs at an almost zero rate, unlike AdaEnetp and AdaEnetpbh which have 35% and 10% FPR respectively. Only Lasso and Enet have nonzero Misclassification Rrates.When N=100 per group, the AdaLassop outperforms all other methods in terms of TPR, FPR and Misclassification Rate, while AdaEnetp and Adaenetpbh suffer higher FPRs.

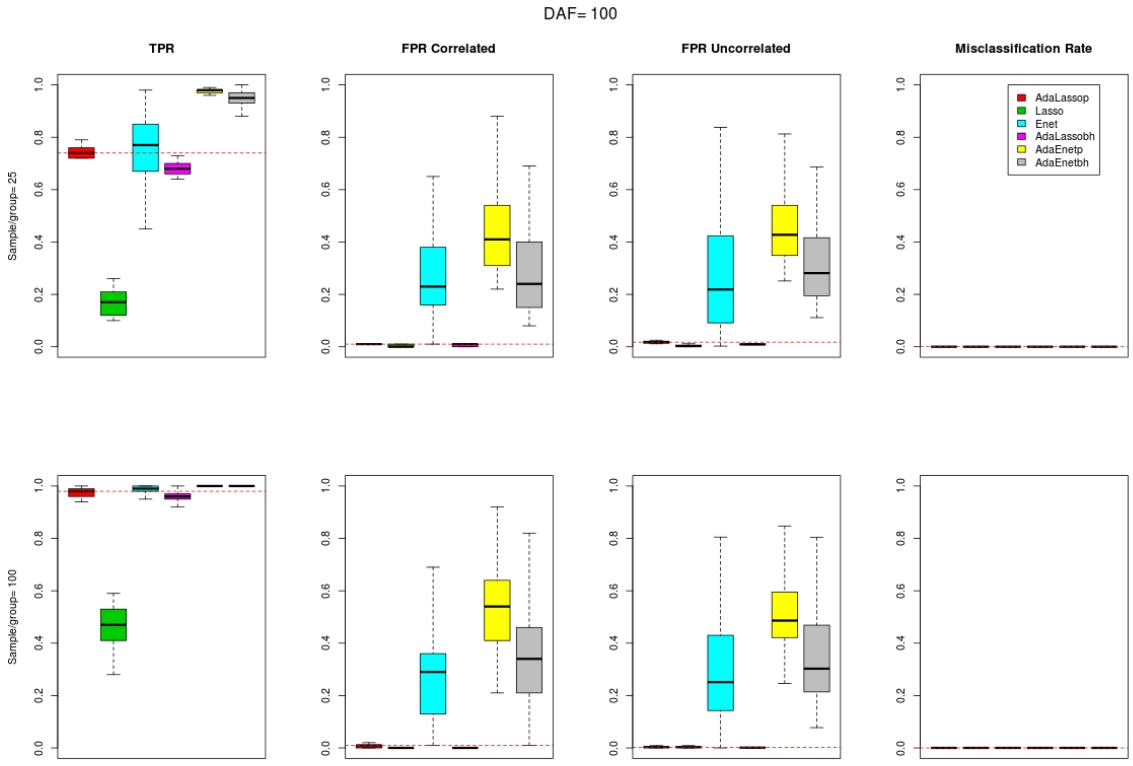

Figure 6.4: Boxplots for Design 2 with 100 DAFs

Figure 3.4 is set up as figure 3.3 but for 100 DAFs. When N=25 per group, the TPR for AdaLassop has average performance and is outperformed by AdaEnetp and AdaEnetpbh by a margin of almost 25%. Furthermore, AdaLassop controls FPR for the correlated non-DAFs at

an almost zero rate, unlike AdaEnetp and AdaEnetpbh which have 40% and 25% FPR. When N=100 per group, the AdaLassop outperforms all other methods in terms of TPR, FPR and Misclassification Rate, while AdaEnetp and Adaenetpbh sustain the same median FPR. As before, the FPR of AdaLassop on correlated non-DAFs with DAFs is almost zero.

6.2.3 DESIGN 3

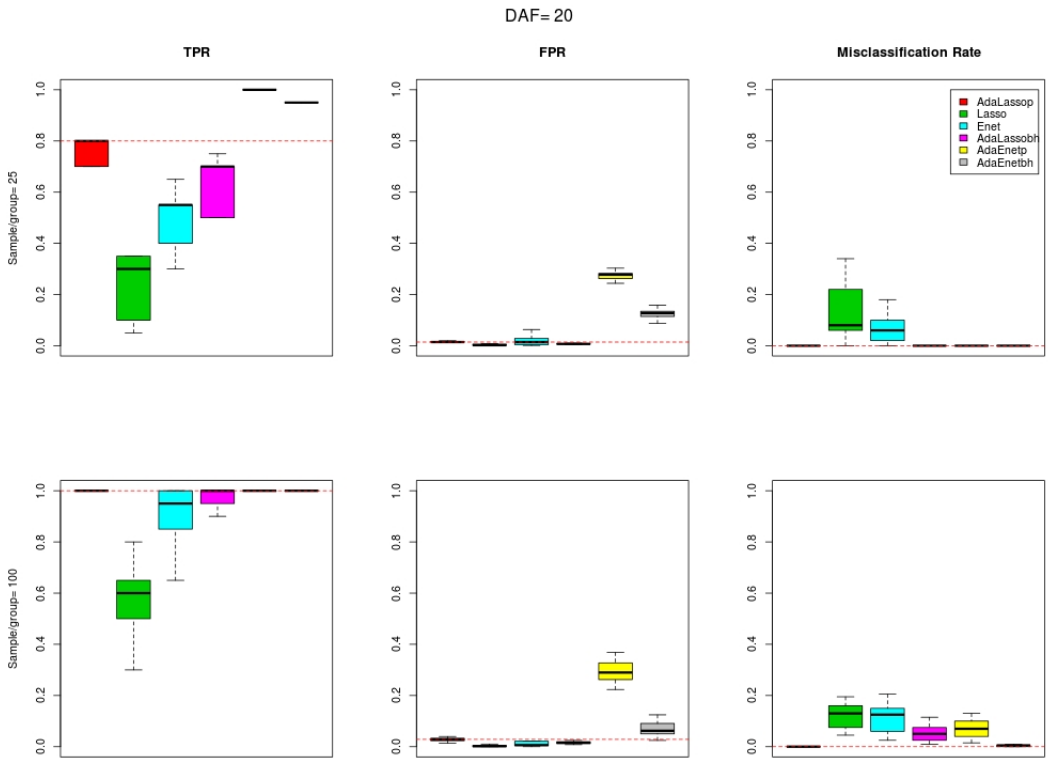

Figure 6.5: Boxplots for Design 3 with 20 DAFs

Figure 3.5 shows True Positive Rates for the correlated DAFs in the first column, False Positive Rates for non-DAFs in the second column and Misclassification Rates in the third column for sample sizes 25 per group on the top row and 100 per group on the bottom row over 25 repetitions for 100 DAFs. The red dotted line indicates the median performance of the AdaLassop for all measures.

The correlation structure in Design 3 is such that DAF 1 is correlated to DAF 11, DAF 2 is correlated to DAF 12 and so on until DAF 10 is correlated to DAF 20 at levels between 0.65 and 0.8 across conditions. When N=25 per group, the TPR for AdaLassop averages around 80% and is outperformed only by AdaEnetp and AdaEnetpbh by a margin of 10%. Further, AdaLassop controls FPR for the non-DAFs at an almost zero rate, unlike AdaEnetp and AdaEnetpbh which have 10% and 25% FPR respectively. Only Lasso and Enet have nonzero Misclassification Rates. When N=100 per group, the AdaLassop outperforms all other methods in terms of TPR and Misclassification Rate while touting an FPR that is almost nonexistent.

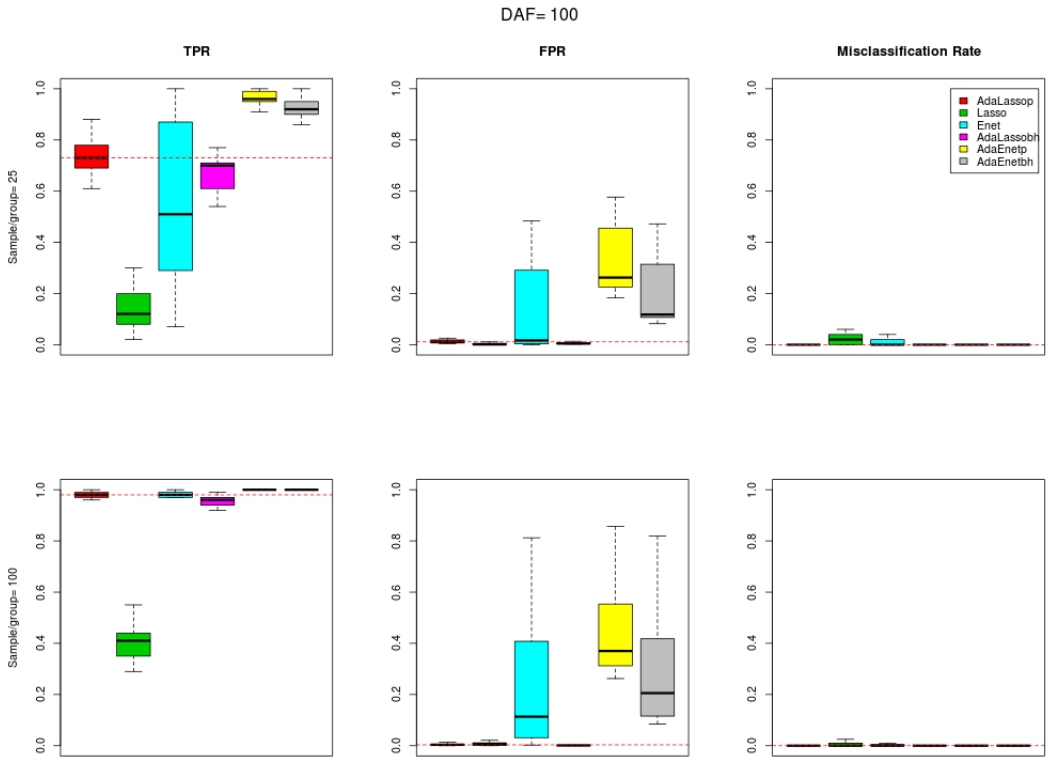

Figure 6.6: Boxplots for Design 3 with 100 DAFs

Figure 3.6 is set up as Figure 3.5 but for 100 DAFs. When N=25 per group, the TPR for AdaLassop averages around 75% and is outperformed only by AdaEnetp and AdaEnetpbh by a margin of 20%. Further, AdaLassop controls FPR for the non-DAFs at an almost zero

rate, unlike AdaEnetp and AdaEnetpbh which have 25% and 10% FPR respectively. Only Lasso and Enet have nonzero Misclassification Rates. When N=100 per group, the AdaLassop outperforms all other methods in terms of TPR and Misclassification Rate while touting an FPR that is almost nonexistent.

In conclusion, AdaLassop has an average but stable TPR performance ranging between 60% to 80% in low dimensions (N=25 per group), exhibits no misclassification in simulations while controling FPR at almost 0% for both correlated non-DAFs with DAFs and purely uncorrelated non-DAFs. In larger sample sizes, AdaLassop exhibits optimal performances in TPR, FPR and Misclassification Rate.

7 Real Data Analysis

7.1 Background

AdaLassop analysis was applied to 50 DNA stool samples collected from male Chinese subjects with type-2 diabetes (N=25) and non-diabetic controls (N=25) in the original dataset (Qin,J. *et al,* 2012) comprising 345 DNA stool samples. First, each sample was aligned against the bacterial reference in NCBI using BLASTN. Then TAEC was executed (Sohn *et al.*, 2014) to estimate the abundance of each feature or bacteria at the species level.

7.2 Data Diagnostics

A brief analysis shows that there is a high degree of correlation between the features in the diabetes dataset (Figure 4.1). Upon closer inspection, more than 80% of the features in the diabetes data are correlated with at least another feature at levels higher than 0.65 and 53% at levels higher than 0.8. This situation is optimal for AdaLassop since it deals very well with variable selection in presence of correlation.

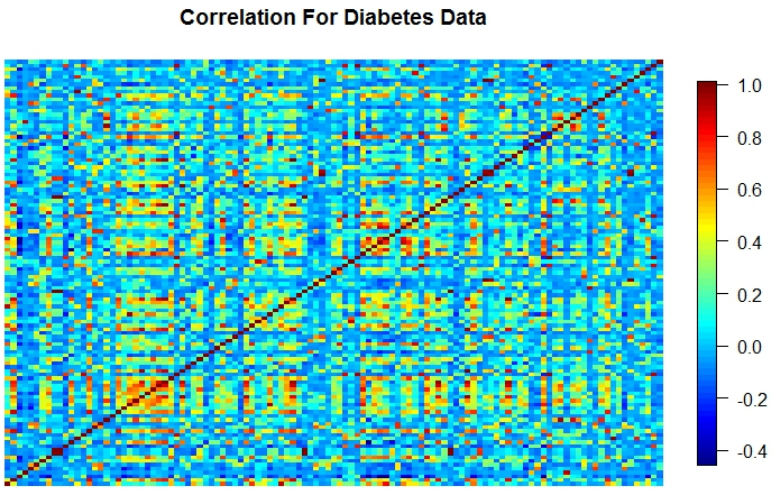

Figure 7.1: Correlation plot for diabetes data

7.3 Analysis

Prior to analysis, we screened the diabetes data so that at most 90% of the data for a given feature per condition comprises zeros. This resulted in a dataset with 182 features. The variable selection procedure was repeated over 25 times on the real data to see the stability of the solutions. Figure 4.2 shows that the performances of all methods other than the Adalassop and AdaLassobh are highly varied. AdaLassop and Adalassobh however consistently selected the same features each round. The Elastic Net exhibits the highest misclassification error, followed by AdaEnetbh, AdaEnetp, Lasso, AdaLassobh and eventually AdaLassop.

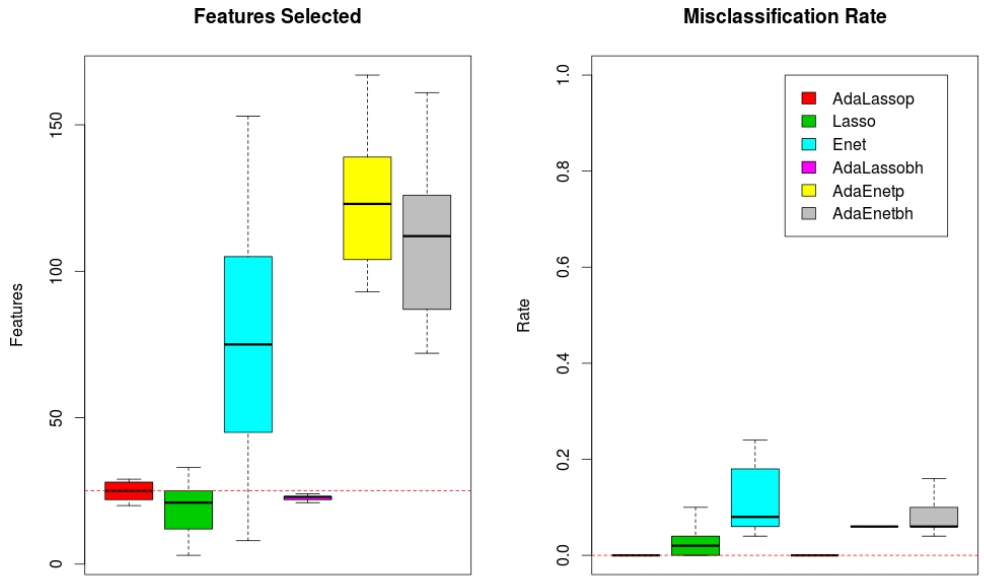

Figure 7.2: Features selected and Misclassification Rates

Features selected from the real dataset and their corresponding coefficients are sorted by magnitude in table 4.1, where the left diagram is a boxplot of the number of features selected at each of the 25 repetitions and the right diagram shows the corresponding Misclassification Rate at each variable selection step.

The bacterial strains *Desulfitobacterium hafniense Y51*, *Fusobacterium nucleatum* and *Bifidobacterium longum* have been documented to be key players in the KEGG pathways for type-2 diabetes. The *Clostridium difficile* is seen increasingly among patients with type-2 diabetes (Hassan SA, *et al.* 2011). Research conducted by Shakov *et. al* in 2013 shows that diabetes is a significant risk factor to recurrences or relapses of Clostridium difficile-associated diarrhea (CDAD). The pathogenic strain produces toxin A and B that cause diarrhea and colitis via colonic mucosal inflammation, which in turn causes subsequent damage.

Symbiobacterium thermophilum is a bacterial strain that thrives in high temperatures and is dependent on co-culture with the *bacillus* strain for growth. Recent research indicates it has been used in site-saturation mutagenesis to synthesize the drug D-Phenylanaline for the treatment of type-2 diabetes (Gao X *et al.*, 2013).

The *Prevotella intermedia 17* strain is an oral human pathogen that causes Periodontitis (Guthmiller, Novak 2002), a condition characterized by the inflammation of the tissues around the teeth often leading to shrinkage of gums and loosening of teeth. Preshaw *et al.* published in 2011 that emerging evidence seems to suggest that there is a mutual relationship between this bacterial strain and diabetes: as diabetes increases the risk of Periodontitis, periondontal inflammation in turn adversely affects glycaemic control, resulting in a viscious cycle.

The *Enterococcus faecalis* strain has been associated with feet infections in diabetic patients (Vinodkumar C S, *et al.* 2011). This bacterial strain occurs naturally in the lower intestinal tract, mouth and vaginal tract on both humans and animals. In healthy individuals, *faecalis* poses no adverse effects to the host. However, this bacterium can in fact induce lethal infections in hosts with impaired immune systems such as those suffering from diabetes. It is especially virulent due to its swift development of drug immunity, making it difficult to control in clinical settings (Zhang *et al.* 2013). At present, research suggests the use of bacteriophages in dealing with the *faecalis* strain.

Name	Coefficients
Prevotella ruminicola 23	-3.508293e-02
[Clostridium] sticklandii	-1.478696e-02
Tannerella forsythia ATCC 43037	-8.282649e-03
Bartonella henselae str. Houston-1	-1.796085e-03
Bacteroides fragilis NCTC 9343	-1.182878e-03
Treponema denticola ATCC 35405	-8.315142e-04
Clostridium tetani E88	-4.173912e-04
Escherichia coli ETEC H10407	-3.844868e-04
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	-2.319692e-04
Desulfovibrio vulgaris str. 'Miyazaki F'	-1.539682e-04
Shigella sonnei Ss046	-1.208177e-04
Shigella dysenteriae Sd197	-3.108853e-06
Citrobacter koseri ATCC BAA-895	1.306721e-06
Clostridium perfringens ATCC 13124	3.165514e-06
Alkaliphilus metalliredigens QYMF	2.049532e-05
Desulfovibrio alaskensis G20	5.306521e-05
Porphyromonas gingivalis W83	5.598878e-05
Enterococcus faecalis V583	1.120734e-04
Prevotella intermedia 17	1.190024e-04
Shigella sonnei 53G	3.062291e-04
Clostridium beijerinckii NCIMB 8052	4.626979e-04
Symbiobacterium thermophilum IAM 14863	5.512616e-04
Clostridium difficile 630	1.936392e-03
Bifidobacterium longum DJO10A	7.072194e-03
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	1.265155e-02
Escherichia coli 042	1.530837e-02
Desulfitobacterium hafniense Y51	1.578469e-02

Table 7.1: List of features selected from the diabetes data with the corresponding coefficients sorted in order of magnitude.

8 Discussion

Now since the p-values are used as weights in the variable selection, one might pose the following question: why not simply use the p-values themselves as indicators for whether a feature is DAF or not? In other words, conduct a Zero Inflated Negative Binomial regression of each individual feature against the phenotype condition? If the calculated p-values from each univariate regression is less than some prespecified significant level α, one would say that the feature is differentially abundant. Simulations indicate that this would indeed yield similar results. However this simply addresses the questions of "which features are differentially abundant?"and "how are they differentially abundant?". AdaLassop does all this, in addition to providing a holistic predictive model that shows the extent of the effects of selected DAFs on diabetes in presence of other selected DAFs.

A close alternative to the AdaLassop is the Group Lasso. The Group Lasso is based on the Lasso whose adaptive weights are a function of the number of covariates in the model and the group size (Yuan and Lin, 2007). This method works very well with correlated features. One of the main drawbacks however is that a QR decomposition must be calculated initially for all groups (Simon and Tibshirani, 2011). Furthermore, some prior insight is necessary to know the group size before executing Group Lasso. AdaLassop on the other hand, does not require information on group sizes and is initiated with a ZINB regression.

One of the benefits of the AdaLassop is that it integrates prior information the researcher has of the data into the modelling process. In addition, this method is more convenient than the original adaptive models since those require the calculation of an optimal γ parameter for the adaptive weights. AdaLassop is a straightforward application of extracting the p-values from the ZINB regression and using the p-values as adaptive weights. This is a novel application of the p-values since there is no need to impose any p-value threshold to identify statistical significance in the model.

That is not to say this method is not without limitations. For instance, AdaLassop works only if the regression in the adaptive weights estimation is not singular. Further, the theoretical

asymptotic behavior of the AdaLassop was not proven, unlike the original Adaptive Lasso. Although, this proposed method works effectively with 2 conditions, one could extend this model to a multinomial Logit model using the (mlogit) package in R.

9 REFERENCES

[1] Achim Zeileis, Christian Kleiber, Simon Jackman (2008). Regression Models for Count Data in R. Journal of Statistical Software 27(8).

[2] Cameron, A. C. and Trivedi, P. K. (1998). Regression Analysis of Count Data. New York: Cambridge Press

[3] Consul, P. C. and G. C. Jain (1970). On the generalization of Poisson distribution. Ann. Math. Statist. 41 (4), 1387.

[4] Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 57, 289-300.

[5] Dudley, R. M. (2003). Mathematical Statistics, 18.466 lecture notes, Spring 2003. On MIT OCW (OpenCourseWare) website, 2004.

[6] Erhardt, V., Czado, C. (2009) A method for approximately sampling high-dimensional count variables with prespecified Pearson correlation.

[7] Fan, J. & Li, R. (2006), 'Statistical challenges with high dimensionality: Feature selection in knowledge discovery', Proceedings of the Madrid International Congress of Mathematicians 2006 Vol. III, 595-622.

[8] Fan J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', Journal of the American Statistical Association 96, 1348-1360.

[9] Gao X, Huang F, Feng J, *et al.* Engineering the meso-Diaminopimelate Dehydrogenase from Symbiobacterium thermophilum by Site Saturation Mutagenesis for d-Phenylalanine Synthesis. Applied and Environmental Microbiology. 2013;79(16):5078-5081. doi:10.1128/AEM.01049-13.

[10] Greene, William (2003) Econometric Analysis. 5th ed. Upper Saddle River, N.J.: Prentice Hall, Print.

[11] Greene, William H, (2008), Functional forms for the negative binomial model for count data, Economics Letters, 99, (3), 585-590

[12] Guthmiller JM, Novak KF. Periodontal Diseases. In: Brogden KA, Guthmiller JM, editors. Polymicrobial Diseases. Washington (DC): ASM Press; 2002. Chapter 8. Available from: http://www.ncbi.nlm.nih.gov/books/NBK2496/

[13] Hassan SA, *et al.* (2011), Hospital-acquired Clostridium difficile infection among patients with type 2 diabetes mellitus in acute medical wards, J R Coll Physicians Edinb. 2013;43(2):103-7. doi: 10.4997/JRCPE.2013.203.

[14] Hu M-C, Pavlicova M, Nunes EV. Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. The American journal of drug and alcohol abuse. 2011;37(5):367-375. doi:10.3109/00952990.2011.597280.

[15] Hausman, Jerry A., and Bronwyn H. Hall. Econometric Models for Count Data with an Application to the Patents-R & D Relationship. Cambridge, Mass.: National Bureau of Economic Research, 1984. Print.

[16] Hilbe, Joseph M. Negative Binomial Regression. Cambridge: Cambridge UP, 2007.

[17] Hsiao, Elaine Y. *et al.* (2013) 1.Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders; Cell , Volume 155 , Issue 7 , 1451 - 1463

[18] Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol., 3, REVIEWS0003.

[19] Huson,D. *et al.* (2009) Methods for comparative metagenomics. BMC Bioinformatics, 10 (Suppl. 1), S12.

[20] N. Ismail, H. Zamani (2013) Estimation of Claim Count Data Using Negative Binomial,

Generalized Poisson, Zero-inflated Negative Binomial and Zero-inflated Generalized Poisson Regression Models Casualty Actuarial Society E-Forum Spring (2013)

[21] Jianqing Fan and Jinchi Lv (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). Journal of Royal Statistical Society B, 70, 849-911.

[22] Johnston, J. Econometric Methods. 2d ed. New York: McGraw-Hill, 1971. Print.

[23] Kunin,V. bibitemnotes (2008) A bioinformatic's guide to metagenomics. Microbiol. Mol. Biol. Rev., 72, 557.

[24] Mullahy J (1986). âĂIJSpecification and Testing of Some Modified Count Data Models.âĂİ Journal of Econometrics, 33, 341âĂŞ365.

[25] N. Simon and R. Tibshirani. Standardization and the group lasso penalty. Technical report, Stanford University, March 2011

[26] Preshaw, P. M. *et al.* âĂIJPeriodontitis and Diabetes: A Two-Way Relationship.âĂİ Diabetologia 55.1 (2012): 21âĂŞ31. PMC. Web. 29 June 2015.

[27] Pookhao N, *et al.* (2015) A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes. Bioinformatics, 31:158-165

[28] Qin,J. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490, 55âĂŞ60.

[29] Rapaport,F. *et al.*. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol., 14, R95.

[30] R. Shakov, R.S. Salazar, S.K. Kagunye, W.J. Baddoura, V.A. DeBari, (2011), Diabetes mellitus as a risk factor for recurrence of Clostridium difficile infection in the acute care hospital setting Am J Infect Control, 39 (2011), pp. 194âĂŞ198

[31] Sohn M, An L, Pookhao N, Li Q. Accurate genome relative abundance estimation for closely related species in a metagenomic sample. BMC Bioinformatics 2014, 15:242 .

[32] Sohn M, Du R and An L. A robust approach for identifying differentially abundant features in metagenomic samples. Bioinformatics, 2015 Mar 19. pii: btv165

[33] Tuenter, H. J. H. (2000), On the generalized Poisson distribution. Statistica Neerlandica, 54: 374âĂŞ376. doi: 10.1111/1467-9574.00147

[34] Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', Journal of the Royal Statistical Society, Series B 58, 267-288.

[35] Vinodkumar C S, *et al.* 2011 Isolation of bacteriophages to multi-drug resistant Enterococci obtained from diabetic foot: A novel antimicrobial agent waiting in the shelf?. Indian J Pathol Microbiol 2011;54:90-5

[36] Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat., 9, 60-62.

[37] Winkelmann, Rainer. Econometric Analysis of Count Data. 5th ed. Berlin: Springer, 2008. Print.

[38] Wooley,J. and Ye,Y. (2010) Metagenomics: facts and artifacts, and computational challenges. J. Comp. Sci. Tech., 25, 71-81.

[39] Yi Yang and Hui Zou (2013). gcdnet: LASSO and (adaptive) Elastic-Net penalized least squares, logistic regression, HHSVM and squared hinge loss SVM using a fast GCD algorithm. R package version 1.0.4. http://CRAN.R-project.org/package=gcdnet

[40] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68 (1):49âĂŞ67, 2007.

[41] Zhang W, Mi Z, Yin X, Fan H, An X, Zhang Z, et al. (2013) Characterization of Enterococcus faecalis Phage IME-EF1 and Its Endolysin. PLoS ONE 8(11): e80435. doi:10.1371/journal.pone.0080435

[42] Zou, H. (2006), 'The adaptive lasso and its oracle properties', Journal of the American Statistical Association 101, 1418-1429.

[43] Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', Journal of the Royal Statistical Society, Series B 67, 301-320.

[44] Zou, Hui; Zhang, Hao Helen. On the adaptive elastic-net with a diverging number of parameters. Ann. Statist. 37 (2009), no. 4, 1733–1751. doi:10.1214/08-AOS625. http://projecteuclid.org/euclid.aos/1245332831.