# Comparison of Classification Methods for Pathogen Detection with High-Dimensional Mi-Fi Data


Huizi Wang

Advisor: Dr. Joshua D Habiger, Department of Statistics, Oklahoma State University
Outside committee: Dr. Kitty Cardwell, Institute of Biosecurity and Microbial Forensics, Department of Entomology and Plant Pathology, Oklahoma State University

# Outline

# Microbe Finder (Mi-Fi)

- Microbe Finder (Mi-Fi) is a diagnostic tool developed by researchers at Oklahoma State University which uses high-throughput sequencing technology to measure the abundance of a pathogen in a sample.
- Data are generated as follows [21]:
  - Step 1: The unique RNA sequence of a pathogen is decomposed into multiple smaller sequences called "e-probes" [9], [8].
  - Step 2: The RNA sequence of a sample to be tested is provided to Mi-Fi platform [3].
  - Step 3: Scan between the RNA sequence of a sample and "e-probes" provided [8].
  - Step 4: The number of "e-probes hits" and the "e-probe scores" in each sample are recorded for each e-probe [9].

# Mi-Fi Data (CLas data set)

| ID | $Y_i$ | $X_{ij}$ | | | | $Z_i$ |
|---|---|---|---|---|---|---|
| | | $X_{i1}$ | $X_{i2}$ | $\cdots$ | $X_{i8842}$ | |
| 1 | 0 | 12 | 13 | $\cdots$ | 0 | 254.1 |
| 2 | 0 | 1 | 1 | $\cdots$ | 0 | 37.88 |
| 3 | 0 | 11 | 19 | $\cdots$ | 0 | 169.93 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 21 | 1 | 2 | 17 | $\cdots$ | 0 | 881.2 |
| 22 | 1 | 3 | 8 | $\cdots$ | 0 | 393.7 |

- $Y_i$: Diagnostic result (qPCR) from lab for sample $i$. 1 if the sample contains a pathogen, and 0 otherwise.
- $X_{ij}$: The number of hits for a sample $i$ in e-probe $j$.
- $Z_i$: Total Score of a sample $i$.
  - Continuous.
  - Incorporates number of hits and quality of each hit.

# Research Problems

- The basic goal is to determine if a sample should be classified as pathogenic or healthy using Mi-Fi data.
- The challenge is that pathogen abundance is measured as the number of "e-probe hits" for at least 10 and up to 10,000 "e-probes", depending on the pathogen of interest, and there are a relatively few number of samples to train a classifier.
- Hence this is a classification problem
  - One dimension classification problem with a single continuous variable "total score" $Z_i$.
  - High-dimension (HD) classification problem with count data $X_{ij}$.

# Simple Models

- There are three simple classifiers base on models Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and logistic regression.
- Denote the conditional probability mass function (PMF) of $Y$ given $Z$ by $p(y \mid z)$ as shorthand for $Pr(Y = y \mid Z = z)$.
- For LDA and QDA models, we state $\hat{Y} = 1$ if $p(1 \mid z) = \frac{\pi_1 f_1(z)}{\pi_1 f_1(z) + \pi_0 f_0(z)} > K$ for some $K$.
  - LDA model,
    - $f_0(z)$: $N(\mu_0, \sigma^2)$
    - $f_1(z)$: $N(\mu_1, \sigma^2)$
  - QDA model,
    - $f_0(z)$: $N(\mu_0, \sigma_1^2)$
    - $f_1(z)$: $N(\mu_1, \sigma_2^2)$
- For logistic regression model, we state $\hat{Y} = 1$ if $\log(\frac{p(1|z)}{p(0|z)}) = \beta_0 + \beta_1 z > K$ for some $K$.

# Bayes Decision Boundary

- Bayes decision boundary (when $p = 1$): solves $p(1 \mid z) = \frac{1}{2}$ for $z$ when $\pi_1 = \pi_0$.
  - LDA: $\hat{Y} = 1$ if $z > \frac{\mu_1 + \mu_0}{2}$
  - QDA: $\hat{Y} = 1$ if $z > \frac{(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}) - \sqrt{\frac{(\mu_0 - \mu_1)^2}{\sigma_0^2 \sigma_1^2} - (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}) \times 2 \log \frac{\sigma_1}{\sigma_0}}}{(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2})}$
  - Logistic regression: $\hat{Y} = 1$ if $z > -\frac{\beta_0}{\beta_1}$
- Remark: All the Bayesian decision boundaries are closed form of above models.

# Estimation

- LDA and QDA
  - Parameters $\pi_0$ and $\pi_1$ can be estimated using a wide variety of techniques. See Chapter 4 of [11] and Chapter 2 of [15] for details.
  - Parameters $\mu_0$, $\mu_1$, $\sigma_0$ and $\sigma_1$ can be estimated using maximum likelihood estimates. See Chapter 11 of [6] and [5] for details.
- Logistic regression
  - Parameters $\beta_0$ and $\beta_1$ can be estimated using standard maximum likelihood estimates.

# Result

- The estimated decision boundary with $K = 0$ and $\pi_0 = \pi_1 = 0.5$ founded by plugging in estimates in the Bayes decision boundary.
  - For LDA, it is $\hat{z} = 2885.04$.
  - For QDA, it is $\hat{z} = 430.34$.
  - For logistic regression, it is $\hat{z} = 339.34$.

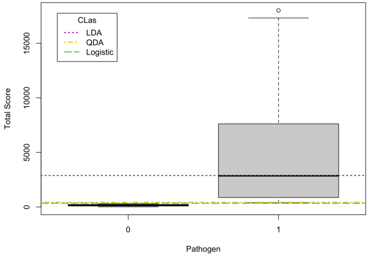



Figure: A Box plot of pathogen vs total score with decision boundaries of $C$Las data set.

# Assessment

- The LOOCV-based confusion matrix with $K = 0$ and $\pi_0 = \pi_1 = 0.5$ of three classifiers for $C$Las data set.

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 5 | 6 | 11 |
| | 0 (N) | 0 | 11 | 11 |
| | Total | 5 | 17 | 22 |

Figure: LDA

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 10 | 1 | 11 |
| | 0 (N) | 0 | 11 | 11 |
| | Total | 10 | 12 | 22 |

Figure: QDA

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 10 | 1 | 11 |
| | 0 (N) | 0 | 11 | 11 |
| | Total | 10 | 12 | 22 |

Figure: Logistic

  - For LDA, FN $= 6$.

# Assessment

- The LOOCV-based summary with $K = 0$ and $\pi_0 = \pi_1 = 0.5$ of the five classification metrics for $C$Las data set.

| Classifier | FPR | TPR | MR | $F_1$ |
|---|---|---|---|---|
| LDA | 0 | 0.45 | 0.27 | 0.62 |
| QDA | 0 | 0.91 | 0.05 | 0.95 |
| Logistic | 0 | 0.91 | 0.05 | 0.95 |

- Useful metrics
  - FPR is Type I Error ($\alpha$).
  - TPR is Power ($1 - \beta$).
  - MR is misclassification rate.
  - $F_1$ is denotes the number of TPs among the mean of predicted positives (precision) and the mean of real positives [16], [22].
  - For many other useful metrics, see [16], [14], [10] and [22] for a detailed review.
- LDA performs poorly because the model of LDA assumes equal variances.

# Assessment

- Receiver operating characteristics (ROC) curve
  - ROC curve plots FPR vs TPR.
  - ROC curve depicts trade-offs between FPR and TPR as we change $K$.
- Area under the ROC curve (AUROC)
  - AUROC summarizes the ROC curve.
  - If the AUROC is 1 then the classifier is "perfect". If the AUROC is 0.5 then the classifier is poor.

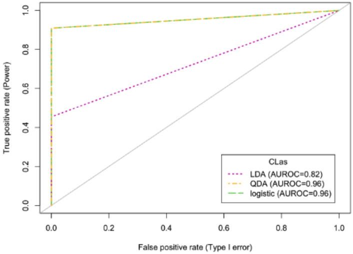



Figure: An ROC graph with three classifiers for $CLas$ data set.

# Notation

- Notation of high-dimensional count data. For short, denote
  - $Pr(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p, Y = y)$ by $p(x_1, x_2, \dots, x_p, y) = p(\boldsymbol{x}, y)$.
  - $Pr(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p \mid Y = y)$ by $p(x_1, x_2, \dots, x_p \mid y) = p(\boldsymbol{x} \mid y)$.
  - $Pr(Y = y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$ by $p(y \mid x_1, x_2, \dots, x_p) = p(y \mid \boldsymbol{x})$.

# Multivariate Approaches

- Assume $p(\boldsymbol{x} \mid y) = \prod_{j=1}^{p} p(x_j \mid y)$. See [2], [7] and [8] for details.
- Specify models for $\boldsymbol{X} \mid y$ such as
  - Poisson [24]
  - Zero-Inflated Poisson (ZIP): considers high proportion of zeros [13].
  - Negative Binomial (NB): considers over-dispersion [24], [18], [1].
  - Zero-Inflated Negative Binomial (ZINB): considers high proportion of zeros and over-dispersion [28].

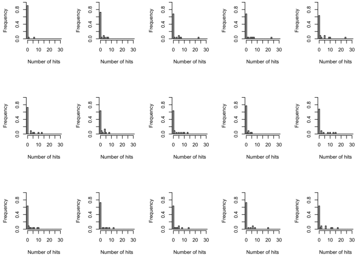



Figure: Zero-inflation plot

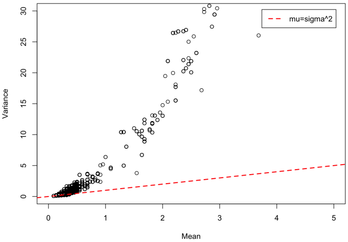



Figure: Over dispersion plot

# Poisson Model and Poisson Linear Discriminant Analysis (PLDA) Classifier

- Poisson model
  - Assume $X_j \mid y \overset{\text{ind}}{\sim} \text{POI}\Big(\mu_j(y)\Big)$ so that $\mu_j(0)$ is mean of $X_j$ when $y = 0$ and $\mu_j(1)$ is mean of $X_j$ when $y = 1$.
- PLDA classifier was proposed by [24].
- **Theorem 1** Observe, the PLDA classifier $\log\left[\frac{p(\boldsymbol{x}|1)}{p(\boldsymbol{x}|0)}\right] > K$ can be written as

$$
\begin{aligned}
C(\boldsymbol{x}) &\propto \log\left[\frac{p(\boldsymbol{x} \mid 1)}{p(\boldsymbol{x} \mid 0)}\right] \\
&= -\sum_{j=1}^{p}\Big(\mu_j(1) - \mu_j(0)\Big) + \sum_{j=1}^{p}\Big[\log\Big(\mu_j(1)\Big) - \log\Big(\mu_j(0)\Big)\Big]x_j \\
&\equiv \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \beta_{1j} x_j .
\end{aligned}
$$

# Zero-Inflated Poisson (ZIP) Model and Zero-Inflated Poisson Logistic Discriminate Analysis (ZIPLDA) Classifier

- ZIP model
  - Assume $X_j \mid y \overset{\text{ind}}{\sim} \begin{cases} \text{POI}\Big(\mu_j(y)\Big), & \text{with probability} \quad 1-\pi_j(y), \\ 0, & \text{with probability} \quad \pi_j(y), \end{cases}$ so that $\mu_j(0)$ is mean of $X_j$ when $y=0$ and $\mu_j(1)$ is mean of $X_j$ when $y=1$. Similarly, $\pi_j(0)$ is probability of $X_j=0$ when $y=0$ and $\pi_j(1)$ is probability of $X_j=0$ when $y=1$.
- ZIPLDA classifier was proposed by [27].

# Zero-Inflated Poisson (ZIP) Model and Zero-Inflated Poisson Logistic Discriminate Analysis (ZIPLDA) Classifier

- **Theorem 2** Observe, the ZIPLDA classifier $\log\left[\frac{p(\boldsymbol{x}|1)}{p(\boldsymbol{x}|0)}\right] > K$ can be written as

$$
\begin{aligned}
C(\boldsymbol{x}) &\propto \log\left[\frac{p(\boldsymbol{x} \mid 1)}{p(\boldsymbol{x} \mid 0)}\right] \\
&= \sum_{j=1}^{p}\left[-\Big(\mu_j(1)-\mu_j(0)\Big)+\log\left(\frac{1-\pi_j(1)}{1-\pi_j(0)}\right)\right]+\sum_{j=1}^{p}\Big[\log\Big(\mu_j(1)\Big)-\log\Big(\mu_j(0)\Big)\Big]x_j \\
&+\sum_{j=1}^{p}\left[\log\left(\frac{\pi_j(1)+\Big(1-\pi_j(1)\Big)e^{-\mu_j(1)}}{\pi_j(0)+\Big(1-\pi_j(0)\Big)e^{-\mu_j(0)}}\right)+\Big(\mu_j(1)-\mu_j(0)\Big)-\log\left(\frac{1-\pi_j(1)}{1-\pi_j(0)}\right)\right]I(x_j=0) \\
&\equiv \boldsymbol{\beta}_0+\sum_{j=1}^{p}\beta_{1j}x_j+\sum_{j=1}^{p}\beta_{2j}I(x_j=0).
\end{aligned}
$$

# Negative Binomial (NB) Model and Negative Binomial Linear Discriminant Analysis (NBLDA) Classifier

- NB model
  - Assume $X_j \mid y \overset{\text{ind}}{\sim} \text{NB}\Big(\mu_j(y), \phi_j\Big)$ so that $\mu_j(0)$ is mean of $X_j$ when $y = 0$ and $\mu_j(1)$ is mean of $X_j$ when $y = 1$, and $\phi_j$ is the dispersion of $X_j$.
- The NBLDA classifier was proposed by [4].
- **Theorem 3** Observe, the NBLDA classifier $\log\left[\frac{p(\boldsymbol{x}|1)}{p(\boldsymbol{x}|0)}\right] > K$ can be written as

$$
\begin{aligned}
C(\boldsymbol{x}) &\propto \log\left[\frac{p(\boldsymbol{x} \mid 1)}{p(\boldsymbol{x} \mid 0)}\right] \\
&= -\sum_{j=1}^{p} \phi_j^{-1} \log\left(\frac{1+\phi_j\mu_j(1)}{1+\phi_j\mu_j(0)}\right) + \sum_{j=1}^{p}\left[\log\Big(\mu_j(1)\Big) - \log\Big(\mu_j(0)\Big) - \log\left(\frac{1+\phi_j\mu_j(1)}{1+\phi_j\mu_j(0)}\right)\right] x_j \\
&\equiv \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \beta_{1j} x_j.
\end{aligned}
$$

# Zero-Inflated Negative Binomial (ZINB) Model and Zero-Inflated Negative Binomial Logistic Discriminate Analysis (ZINBLDA) Classifier

- ZINB model
  - Assume $X_j \mid y \overset{\text{ind}}{\sim} \begin{cases} \text{NB}\Big(\mu_j(y), \phi_j\Big), & \text{with probability} \quad 1-\pi_j(y), \\ 0, & \text{with probability} \quad \pi_j(y), \end{cases}$ so that $\mu_j(0)$ is mean of $X_j$ when $y=0$ and $\mu_j(1)$ is mean of $X_j$ when $y=1$. Similarly, $\pi_j(0)$ is probability of $X_j=0$ when $y=0$ and $\pi_j(1)$ is probability of $X_j=0$ when $y=1$.
- The ZINBLDA classifier was proposed by [28].

# Zero-Inflated Negative Binomial (ZINB) Model and Zero-Inflated Negative Binomial Logistic Discriminate Analysis (ZINBLDA) Classifier

- **Theorem 4** Observe, the ZINBLDA classifier $\log\left[\frac{p(\boldsymbol{x}|1)}{p(\boldsymbol{x}|0)}\right] > K$ can be written as

$$
\begin{aligned}
C(\boldsymbol{x}) &\propto \log\left[\frac{p(\boldsymbol{x}\mid 1)}{p(\boldsymbol{x}\mid 0)}\right]\\
&= \sum_{j=1}^{p}\left[\log\left(\frac{1-\pi_j(1)}{1-\pi_j(0)}\right) - \phi_j^{-1}\log\left(\frac{1+\phi_j\mu_j(1)}{1+\phi_j\mu_j(0)}\right)\right] + \sum_{j=1}^{p}\left[\log\Big(\mu_j(1)\Big) - \log\Big(\mu_j(0)\Big) - \log\left(\frac{1+\phi_j\mu_j(1)}{1+\phi_j\mu_j(0)}\right)\right]x_j\\
&+ \sum_{j=1}^{p}\left[\log\left(\frac{\pi_j(1)+\Big(1-\pi_j(1)\Big)\left(\frac{1}{1+\phi_j\mu_j(1)}\right)^{\phi_j^{-1}}}{\pi_j(0)+\Big(1-\pi_j(0)\Big)\left(\frac{1}{1+\phi_j\mu_j(0)}\right)^{\phi_j^{-1}}}\right) - \log\left(\frac{1-\pi_j(1)}{1-\pi_j(0)}\right) + \phi_j^{-1}\log\left(\frac{1+\phi_j\mu_j(1)}{1+\phi_j\mu_j(0)}\right)\right] I(x_j=0)\\
&\equiv \boldsymbol{\beta}_0 + \sum_{j=1}^{p}\beta_{1j}x_j + \sum_{j=1}^{p}\beta_{2j}I(x_j=0).
\end{aligned}
$$

# Multiple Logistic Regression Approach (MLR)

- Specify the model $Y \mid \boldsymbol{x} \sim \mathrm{Bern}(p(1 \mid \boldsymbol{x}))$ where

$$C(\boldsymbol{x}) \propto \log\left(\frac{p(1 \mid \boldsymbol{x})}{p(0 \mid \boldsymbol{x})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_{1j} x_j + \sum_{j=1}^{p} \beta_{2j} I(x_j = 0). \tag{1}$$

# Summary of Models

- Models
  - Multivariate
    - PLDA
    - ZIPLDA
    - NBLDA
    - ZINBLDA
  - Multiple Logistic Regression
- Main Point
  - **Main Theory:** All classifiers of the form $\hat{Y} = 1$ if

$$C(\boldsymbol{x}) \propto \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \beta_{1j} x_j + \sum_{j=1}^{p} \beta_{2j} I(x_j = 0) > K. \tag{2}$$

  - **Theorem 1-4**
    - $\beta_0$, $\beta_{1j}$ and $\beta_{2j}$ are function of $\mu_j(y)$'s, $\pi_j(y)$'s and $\phi_j$.
    - How to estimate $\beta_0$, $\beta_{1j}$ and $\beta_{2j}$.

# PLDA-based Estimation

- [24] considered unpenalized MLE-type estimators for $\mu_j(y)$'s
  - can be plugged into Theorem 1 expression.
- [24] proposed penalized/shrinkage maximum likelihood estimators for $\mu_j(y)$'s (or a parameter proportion to $d_j(y)$'s)
  - $L_1$ type penalty for tuning parameter $\rho$.
  - $\rho$ is non-negative tuning parameter that can be chosen by cross-validation.
  - When $\rho = 0$, no shrinkage occurs. When $\rho$ is large, many $\hat{\beta}_{1j}$s are 0.
- Remark: [24] is the only one who considered variable selection or $L_1$ type penalized estimators for parameters $\mu_j(y)$'s (or a parameter proportion to $\mu_j(y)$'s) in the PLDA classifier.

# Other Multivariate Approaches Estimation

- Many other multivariate-type estimators of $\mu_j(y)$'s, $\pi_j(y)$'s and $\phi_j$ considered with some type of penalty. For example,
  - [19] proposed Method of Moment Estimator of dispersion parameter $\phi_j$.
  - [25] introduced the generalized shrinkage estimator for dispersion parameter $\phi_j$. This generalized shrinkage estimator shrinks the dispersion parameter $\phi_j$ toward a target value as tuning parameter increases.
  - [17] proposed the $L_2$ type penalized maximum likelihood estimator of $\mu_j(y)$'s, $\pi_j(y)$'s and $\phi_j$.

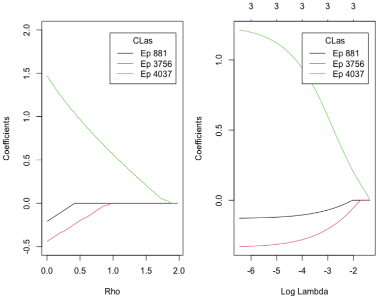
# MLR-based Estimation

- The negative log likelihood is
$\ell(\beta_0, \beta) = -\frac{1}{n}\sum_{i=1}^{n}\Big[y_i \log p(1 \mid \boldsymbol{x}_i) + (1 - y_i)\log p(0 \mid \boldsymbol{x}_i)\Big] =$
$-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\beta_0 + \beta^T \boldsymbol{x}_i\right) - \log\left(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}\right)\right].$
- Common penalized likelihood estimators
  - Ridge Estimator [12] it the solution of $\underset{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{minimize}}\left\{\ell(\beta_0, \beta) + \lambda\|\beta\|_2^2\right\}.$
  - Lasso Estimator [23] is the solution of $\underset{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{minimize}}\left\{\ell(\beta_0, \beta) + \lambda\|\beta\|_1\right\}.$
  - Elastic-net Estimator [29] is the solution of
$\underset{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{minimize}}\left\{\ell(\beta_0, \beta) + \lambda\left[\frac{1}{2}(1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right]\right\}.$
- $\beta$ could be $\boldsymbol{\beta_1}$ in PLDA and NBLDA, or could be $\boldsymbol{\beta_1}$ in ZIPLDA and ZINBLDA by taking $\boldsymbol{\beta_2} = \mathbf{0}$.

- Un-penalized estimates for $\beta_{1j}$ when $p = 3$

| Estimates | E-probe 881 | E-probe 3756 | E-probe 4037 |
|---|---|---|---|
| PLDA | -0.21 | -0.44 | 1.47 |
| MLR | -0.13 | -0.34 | 1.24 |

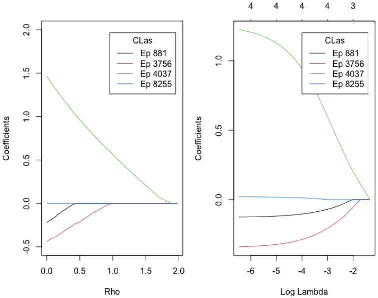
- $L_1$ type penalized estimates for $\beta_{1j}$ when $p = 3$
  - Left: PLDA with tuning parameters $\rho \in [0, 2]$.
  - Right: MLR with tuning parameter $\lambda \in [0, 0.26]$ and $\alpha = 1$.



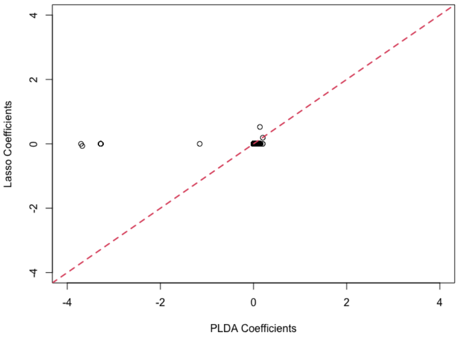
- Un-penalized estimates for $\beta_{1j}$ when $p = 4$

| Estimates | Ep 881 | Ep 3756 | Ep 4037 | Ep 8255 |
|---|---|---|---|---|
| PLDA | -0.22 | -0.44 | 1.47 | 0.01 |
| MLR | -0.13 | -0.35 | 1.26 | 0.02 |

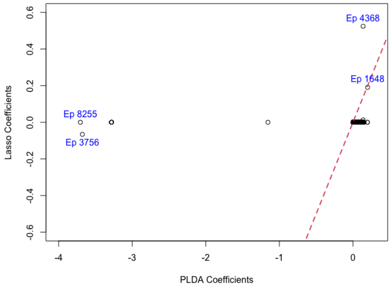
- $L_1$ type penalized estimates for $\beta_{1j}$ when $p = 4$
  - Left: PLDA with tuning parameters $\rho \in [0, 2]$.
  - Right: MLR with tuning parameter $\lambda \in [0, 0.26]$ and $\alpha = 1$.

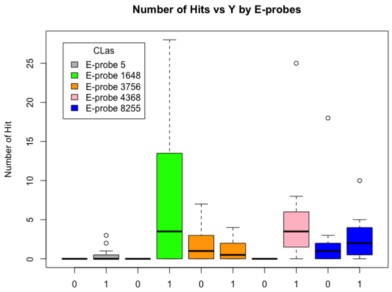


- $L_1$ type penalized estimates for $\beta_{1j}$ when $p = 8842$
  - PLDA
    - Chose LOOCV-based best $\rho$ which has the smallest MR.
    - Using the best $\rho = 0.02$ to fit penalized PLDA.
  - MLR
    - Fix $\alpha = 1$, chose LOOCV-based best $\lambda$ which has the smallest MR.
    - Using the best $\lambda = 0.17$ to fit penalized MLR.



- Many $\hat{\beta}_{1j}$'s have same value 0 between PLDA and MLR $L_1$ type penalized estimates. For example, e-probe 1, 5, 660, 1016, 1341 are shrinkage to 0.
  - Because most of these e-probes don't have any hit when $Y = 0$, and have less than 5 hits when $Y = 1$ for some samples.
- Both PLDA and MLR $L_1$ type penalized estimates are not 0, but it has the same estimates between PLDA and MLR. For example, e-probe 1648.
  - Because this e-probe don't have any hit when $Y = 0$. However, it has many large number of hits when $Y = 1$ for some samples.



- PLDA and MLR $L_1$ type penalized estimates are different. For example, e-probe 4368.
  - PLDA estimates is around 0 but MLR estimates is around 0.5.
  - Because this e-probe don't have any hit when $Y = 0$. However, it has many hits and there is a sample that has a large number of hits when $Y = 1$.
- PLDA and MLR $L_1$ type penalized estimates are different. For example, e-probe 8255.
  - PLDA estimates is around -4 but MLR estimates is 0.
  - Because this e-probe has 18 hits when $Y = 0$ on a sample. However, it has few hits (around 3) when $Y = 1$ for some samples.
  - PLDA thinks this e-probe is important to be included in the classification, but MLR thinks it is unnecessary to be included in the classification. The reason it that PLDA estimates considers difference $\log(\mu_j(1)) - \log(\mu_j(0))$ (or a parameter proportion to $\mu_j(y)$). However, MLR estimates $\beta_{1j}$ using maximum likelihood directly.

| Y | Ep 5 | Ep 1648 | Ep 3756 | Ep 4368 | 8255 |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 4 | 0 | 3 |
| 0 | 0 | 0 | 3 | 0 | 1 |
| 0 | 0 | 0 | 7 | 0 | 2 |
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 3 | 0 | 18 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 10 |
| 1 | 2 | 3 | 2 | 7 | 4 |
| 1 | 0 | 18 | 0 | 8 | 0 |
| 1 | 1 | 10 | 2 | 5 | 2 |
| 1 | 0 | 17 | 4 | 4 | 4 |
| 1 | 3 | 28 | 2 | 25 | 5 |
| 1 | 0 | 4 | 1 | 3 | 1 |
| 1 | 0 | 8 | 0 | 1 | 2 |
| 1 | 0 | 0 | 0 | 2 | 0 |
| 1 | 0 | 1 | 0 | 0 | 3 |
| 1 | 0 | 0 | 0 | 2 | 2 |
| 1 | 0 | 0 | 4 | 5 | 0 |

# Assessment

- MLR approach: LOOCV-based confusion matrix with $K = 0$, $\pi_0 = \pi_1 = 0.5$ of Ridge, Elastic-net and Lasso classifiers with LOOCV-based best tuning parameter $\lambda$ which has smallest MR for $C$Las data set.

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 8 | 4 | 12 |
| | 0 (N) | 0 | 10 | 10 |
| | Total | 8 | 14 | 22 |

Figure: Ridge with $\alpha = 0$ and $\lambda = 3.1$

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 6 | 6 | 12 |
| | 0 (N) | 1 | 9 | 10 |
| | Total | 7 | 15 | 22 |

Figure: Elastic-net with $\alpha = 0.3$ and $\lambda = 0.4$

| | | $\hat{Y}$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| $Y$ | 1 (P) | 6 | 6 | 12 |
| | 0 (N) | 1 | 9 | 10 |
| | Total | 7 | 15 | 22 |

Figure: Elastic-net with $\alpha = 0.5$ and $\lambda = 0.3$

# Assessment

- Multivariate approach: LOOCV-based confusion matrix with $K = 0$ of classifier with LOOCV-based best tuning parameter $\rho$ which has smallest MR for $C$Las data set.

| | | $\hat{Y}$ 1 | $\hat{Y}$ 0 | Total |
|---|---|---|---|---|
| $Y$ | 1 (P) | 6 | 6 | 12 |
| | 0 (N) | 1 | 9 | 10 |
| | Total | 7 | 15 | 22 |

Figure: Elastic-net with $\alpha = 0.7$ and $\lambda = 0.2$

| | | $\hat{Y}$ 1 | $\hat{Y}$ 0 | Total |
|---|---|---|---|---|
| $Y$ | 1 (P) | 6 | 6 | 12 |
| | 0 (N) | 1 | 9 | 10 |
| | Total | 7 | 15 | 22 |

Figure: Lasso with $\alpha = 1$ and $\lambda = 0.17$

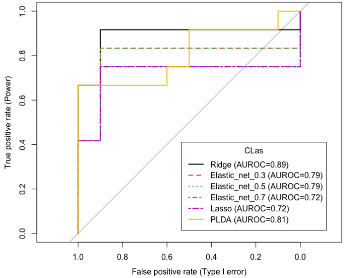
| | | $\hat{Y}$ 1 | $\hat{Y}$ 0 | Total |
|---|---|---|---|---|
| $Y$ | 1 (P) | 8 | 4 | 12 |
| | 0 (N) | 0 | 10 | 10 |
| | Total | 8 | 14 | 22 |

Figure: PLDA with $\rho = 0.02$

# Assessment

- MLR approach: LOOCV-based summary of classification metrics with $K = 0$, $\pi_0 = \pi_1 = 0.5$ of Ridge, Elastic-net and Lasso classifiers with LOOCV-based best tuning parameter $\lambda$ which has smallest MR for $C$Las data set.
- Multivariate approach: LOOCV-based summary of classification metrics with $K = 0$ of classifier with LOOCV-based best tuning parameter $\rho$ which has smallest MR for $C$Las data set.

| Classifier | FPR | TPR | MR | $F_1$ | AUROC |
|---|---|---|---|---|---|
| Ridge | 0 | 0.67 | 0.18 | 0.8 | 0.89 |
| Elastic-net ($\alpha = 0.3$) | 0.1 | 0.5 | 0.32 | 0.63 | 0.79 |
| Elastic-net ($\alpha = 0.5$) | 0.1 | 0.5 | 0.32 | 0.63 | 0.79 |
| Elastic-net ($\alpha = 0.7$) | 0.1 | 0.5 | 0.32 | 0.63 | 0.72 |
| Lasso | 0.1 | 0.5 | 0.32 | 0.63 | 0.72 |
| PLDA | 0 | 0.67 | 0.18 | 0.8 | 0.81 |



# Future Work

- Application for Theorem 2 that incorporate zero-inflation.
  - The negative log likelihood is
    $\ell(\beta_0, \beta) = -\frac{1}{n} \sum_{i=1}^{n} \Big[ y_i \log p(1 \mid \boldsymbol{x}_i) + (1 - y_i) \log p(0 \mid \boldsymbol{x}_i) \Big] = -\frac{1}{n} \sum_{i=1}^{n} \Big[ y_i \Big( \beta_0 + \sum_{g=1}^{G} \beta_g^T \boldsymbol{x}_{i,g} \Big) - \log \Big( 1 + e^{\beta_0 + \sum_{g=1}^{G} \beta_g^T \boldsymbol{x}_{i,g}} \Big) \Big]$.
  - Common penalized likelihood estimators
    - Group Lasso Estimator [26] is the solution of
      $\underset{\beta_0 \in \mathbb{R}, \beta_g \in \mathbb{R}^{p_g}}{\text{minimize}} \Big\{ \ell(\beta_0, \beta) + \lambda \sum_{g=1}^{G} \sqrt{p_g} \|\beta_g\|_2 \Big\}$.
    - Sparse Group Lasso Estimator [20] is the solution of
      $\underset{\beta_0 \in \mathbb{R}, \beta_g \in \mathbb{R}^{p_g}}{\text{minimize}} \Big\{ \ell(\beta_0, \beta) + \lambda \sum_{g=1}^{G} \Big[ (1 - \alpha) \|\beta_g\|_2 + \alpha \|\beta_g\|_1 \Big] \Big\}$.
- Application for Theorem 3 that incorporate over-dispersion.
- Application for Theorem 4 that incorporate zero-inflation and over-dispersion.
- Simulation Study.

# Acknowledgment


- I would like to express my deepest gratitude to my advisor Dr. Joshua D Habiger.
- I would like to express my deepest appreciation to my committee Dr. Lan Zhu and Dr. Ye Liang, and outside committee Dr. Kitty Cardwell.
- Special thanks to Dr. Pratyaydipta Rudra, Dr. Andres Espindola Camacho, and Dami.
- I would like to acknowledge my parents.

[1] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Nature Precedings* (2010), pp. 1–1.

[2] Peter J Bickel and Elizaveta Levina. "Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations". In: *Bernoulli* 10.6 (2004), pp. 989–1010.

[3] Somnath Datta and Dan Nettleton. "Statistical analysis of next generation sequencing data". In: (2014).

[4] Kai Dong et al. "NBLDA: negative binomial linear discriminant analysis for RNA-Seq data". In: *BMC bioinformatics* 17 (2016), pp. 1–10.

[5] Qian Du. "Modified Fisher's linear discriminant analysis for hyperspectral imagery". In: *IEEE geoscience and remote sensing letters* 4.4 (2007), pp. 503–507.

[6] Richard O Duda. *Patter Classification an Scene Analysis*. John Wiley, 1973.

[7] Sandrine Dudoit, Jane Fridlyand and Terence P Speed. "Comparison of discrimination methods for the classification of tumors using gene expression data". In: *Journal of the American statistical association* 97.457 (2002), pp. 77–87.

[8] Andres S Espindola and Kitty F Cardwell. "Microbe finder (mifi®): Implementation of an interactive pathogen detection tool in metagenomic sequence data". In: *Plants* 10.2 (2021), p. 250.

[9] Andres S Espindola et al. "A step towards validation of high-throughput sequencing for the identification of plant pathogenic oomycetes". In: *Phytopathology®* 112.9 (2022), pp. 1859–1866.

[10] Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis. "On clustering validation techniques". In: *Journal of intelligent information systems* 17 (2001), pp. 107–145.

[11] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[12] A Hoerl and R Kennard. *Ridge regression, in 'encyclopedia of statistical sciences', vol. 8*. 1988.

[13] Diane Lambert. "Zero-inflated Poisson regression, with an application to defects in manufacturing". In: *Technometrics* 34.1 (1992), pp. 1–14.

[14] Antonio Maratea, Alfredo Petrosino and Mario Manzo. "Adjusted F-measure and kernel scaling for imbalanced data learning". In: *Information Sciences* 257 (2014), pp. 331–341.

[15] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.

[16] David MW Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020).

[17] Davide Risso et al. "A general and flexible method for signal extraction from single-cell RNA-seq data". In: *Nature communications* 9.1 (2018), p. 284.

[18] Mark D Robinson, Davis J McCarthy and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1 (2010), pp. 139–140.

[19] Krishna Saha and Sudhir Paul. "Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter". In: *Biometrics* 61.1 (2005), pp. 179–185.

[20] Noah Simon et al. "A sparse-group lasso". In: *Journal of computational and graphical statistics* 22.2 (2013), pp. 231–245.

[21] Anthony H Stobbe et al. "E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics". In: *Journal of microbiological methods* 94.3 (2013), pp. 356–366.

[22] Alaa Tharwat. "Classification assessment methods". In: *Applied computing and informatics* 17.1 (2020), pp. 168–192.

[23] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[24] Daniela M Witten. "Classification and clustering of sequencing data using a Poisson model". In: (2011).

[25] Danni Yu, Wolfgang Huber and Olga Vitek. "Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size". In: *Bioinformatics* 29.10 (2013), pp. 1275–1282.

[26] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.1 (2006), pp. 49–67.

[27] Yan Zhou et al. "Classifying next-generation sequencing data using a zero-inflated Poisson model". In: *Bioinformatics* 34.8 (2018), pp. 1329–1335.

[28] Jiadi Zhu et al. "Selecting classification methods for small samples of next-generation sequencing data". In: *Frontiers in Genetics* 12 (2021), p. 642227.

[29] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.