Sparse Learning for Assessing the Association Between Gut Microbiome and Parkinson's Disease

Longhai Li and Man Chen

Department of Mathematics and Statistics
University of Saskatchewan

July 13, 2025
The 3rd JCSDS, Hangzhou, China

Outline

Section 1

Parkinson's Disease and Gut Microbiome

Parkinson's Disease

- PD is the second most common neurodegenerative disease after Alzheimer's disease (AD), with a prevalence of approximately 0.5–1% among those 65–69 years of age, rising to 1–3% among persons 80 years of age and older. With an aging population, both the prevalence and incidence of PD are expected to increase by more than 30% by 2030, which will result in both direct and indirect costs on both society and the economy as a whole.
- In Parkinson's disease, certain nerve cells called neurons in the brain gradually break down or die. Many of the symptoms of Parkinson's are due to a loss of neurons that produce a chemical messenger in your brain called dopamine. When dopamine levels decrease, it causes irregular brain activity, leading to problems with movement and other symptoms of Parkinson's disease.

Causes of Parkinson's Disease

- The cause of Parkinson's disease is unknown, but several factors appear to play a role, including:
  - Genes. Researchers have identified specific genetic changes that can cause Parkinson's disease. But these are uncommon except in rare cases with many family members affected by Parkinson's disease. However, certain gene variations appear to increase the risk of Parkinson's disease but with a relatively small risk of Parkinson's disease for each of these genetic markers.
  - Environmental triggers. Exposure to certain toxins or environmental factors may increase the risk of later Parkinson's disease, but the risk is small.
- Researchers also have noted that many changes occur in the brains of people with Parkinson's disease. These changes include:
  - The presence of Lewy bodies. Clumps of specific substances within brain cells are microscopic markers of Parkinson's disease. These are called Lewy bodies, and researchers believe these Lewy bodies hold an important clue to the cause of Parkinson's disease.

Causes of Parkinson's Disease (continued)

- α-synuclein found within Lewy bodies. Although many substances are found within Lewy bodies, scientists believe that an important one is the natural and widespread protein called alpha-synuclein, also called α-synuclein. It's found in all Lewy bodies in a clumped form that cells can't break down. This is currently an important focus among Parkinson's disease researchers. Researchers have found the clumped alpha-synuclein protein in the spinal fluid of people who later develop Parkinson's disease.

However, it's not clear why these changes occur.

Altered Gut Microbiome Contributes to PD?

- Gastrointestinal (GI) symptoms, including constipation, often precede the motor signs of PD. Lewy bodies and a-synuclein, which are the neuropathological hallmarks of PD, may appear in the gut before they appear in the brain. Colonic inflammation has also been documented in PD.
- Research studies in animals have shown that an altered microbiome might contribute to PD pathology.
  - One study published in Cell showed that there was more α-synuclein accumulation in the brain of the mice with a normal microbiome as compared to the same mice who were raised in a germ-free environment with no bacteria in their gut. This supports the theory that abnormal alpha-synuclein accumulation in the brain is enhanced by a particular microbiome in the gut. See this paper:

    Sampson, T.R., et al. (2016). Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. Cell 167, 1469-1480.e12. 10.1016/j.cell.2016.11.018.

Altered Gut Microbiome Contributes to PD? (continued)

- Other studies showed that transplantation of fecal material from PD mice to normal mice, thereby introducing a "PD microbiome" into mice without PD pathology in their brain, led to impairment of motor function and a decline in brain dopamine. These studies also support the theory that a particular microbiome might be integral in causing PD pathology in the brain. See this paper:

  Matheoud, D., Cannon, T., Voisin, A., Penttinen, A.-M., Ramet, L., Fahmy, A.M., Ducrot, C., Laplante, A., Bourque, M.-J., Zhu, L., et al. (2019). Intestinal infection triggers Parkinson's disease-like symptoms in Pink1$^{-/-}$ mice. Nature 571, 565–569. 10.1038/s41586-019-1405-y.

A Dataset Linking PD and Gut Microbiome

- The original paper that released the dataset:
  Hill-Burns, E.M. et al. (2017). Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Movement Disorders 32, 739–749. 10.1002/mds.26942.
- Dataset: 212 PD cases and 136 control were chosen from the participants enrolled in the Neuro Genetics Research Consortium in Seattle, Washington; Atlanta, Georgia; and Albany, New York. The final dataset included 327 participants, with 197 PD cases and 130 controls.
- URL to retrieve the dataset:
  `https://www.ebi.ac.uk/ena/data/view/PRJEB14674`

QIIME to Preprocess Raw data

- QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data.
- OTU: An operational taxonomic unit (OTU) is an operational definition used to classify groups of closely related sequences.
- Operational taxonomic units (OTUs) are picked based on the closed-reference option, by using the SortMeRNA method against the Green genes 16S rRNA gene sequence database released in August 2013.
- We choose 94% similarity as the similarity threshold, so the output OTUs are at the genus level.

OTU Table

- i is the index of samples, and j is the index of OTUs.
- $\text{OTU}_i^{(j)}$ is the count of fragment sequences of sample i that belong to $\text{OTU}^{(j)}$.

	$\text{OTU}^{(1)}$	$\text{OTU}^{(2)}$	.....	$\text{OTU}^{(q)}$	Age	Sex	Parkinson	Total reads
Sample_1	$\text{OTU}_1^{(1)}$	$\text{OTU}_1^{(2)}$	.....	$\text{OTU}_1^{(q)}$	age_1	sex_1	Y_1	$T_1 = \sum_{j=1}^{q} \text{OTU}_1^{(j)}$
.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....
Sample_n	$\text{OTU}_n^{(1)}$	$\text{OTU}_n^{(2)}$	.....	$\text{OTU}_n^{(q)}$	age_n	sex_n	Y_n	$T_n = \sum_{j=1}^{q} \text{OTU}_n^{(j)}$

Section 2

Transformation of OTU Data

Transformation of OTU Data

In microbiome analysis, people always assume that the phenotype can influence the relative abundance of OTUs, rather than original counts of OTUs. Therefore, when we fit models to OTU dataset, a reasonable transformation for OTU data is necessary, changing count number into information of relative abundance. Here, we choose this transformation of the proportion:

$$\begin{aligned} \widetilde{X}_i^{(j)} &= \log\left(\frac{X_i^{(j)}+1}{\sum_{j=1}^{p} X_i^{(j)}+p}\right) \\ &= \log(X_i^{(j)}+1) - \log(\sum_{j=1}^{p} X_i^{(j)}+p). \end{aligned}$$

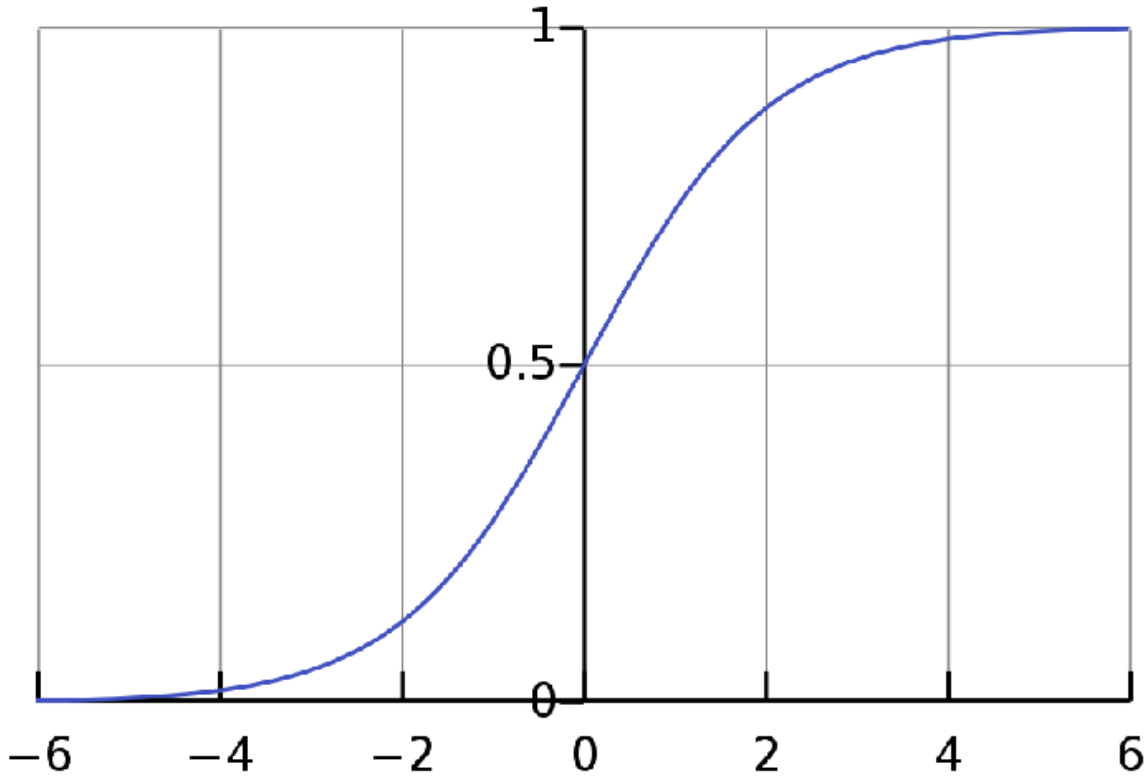
Section 3

Sparse Learning with Logistic Regression and Support Vector Machine (SVM)

Subsection 1

Logistic Regression

Sigmoid Function



- $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}, S \in (0,1).$

Logistic Regression

- $y_i \in \{0, 1\}$, indicator of Parkinson's disease
- $X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)})^T$: OTU features, age and sex, etc.
- A probabilistic model for y_i given X_i:

$$P(y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+....\beta_p x_i^{(p)}}}{1 + e^{\beta_0+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+....\beta_p x_i^{(p)}}}$$

- Another way to express the above model:

$$\log\left(\frac{P(y_i = 1|X_i)}{1 - P(y_i = 1|X_i)}\right) = \beta_0 + \beta^T X_i$$

- Loss Function (minus log-likelihood):

$$J(\beta) = -\sum_{i=1}^{n}\left(y_i(\beta_0 + \sum_{j=1}^{p}\beta_j x_i^{(j)}) - \log(1 + e^{\beta_0+\sum_{j=1}^{p}\beta_j x_i^{(j)}})\right)$$

Subsection 2

Support Vector Machine

Support Vector Machine for Linearly Separable Cases

- $y_i \in \{-1, 1\}$, indicator of Parkinson's disease
- $X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)})^T$: OTU features, age and sex, etc.
- The simplest idea of SVM is to find a hyperplane represented by w and b such that

$$\begin{aligned} \mathbf{w}_i^T x + b \leq 0, \quad & \text{if } y_i = -1 \\ \mathbf{w}_i^T x + b > 0, \quad & \text{if } y_i = 1 \end{aligned}$$

Support Vector Machine for Linearly Separable Cases (continued)

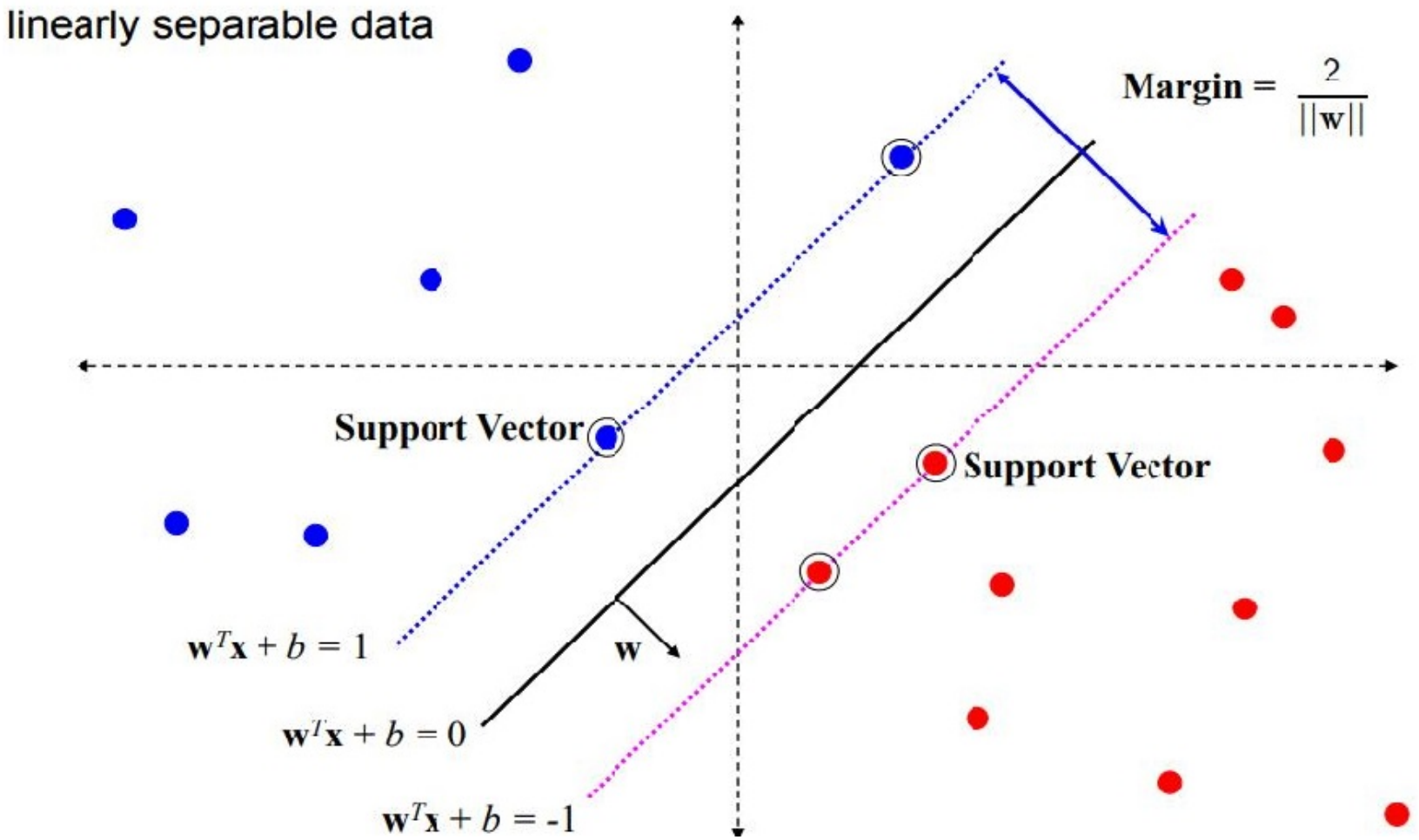


Figure 1: SVM for Linearly Separable Cases

Optimization Formulation for Support Vector Machine

- Based on the idea of separating data points as accurately as possible with a maximization of the margin, SVM is formulated as:

$$\begin{aligned} &\underset{w,b}{\text{maximizing}} \ \frac{2}{\|w\|}, \\ &\text{subject to } y_i(w^T X_i + b) > 1, i = 1, 2, ..., n. \end{aligned} \tag{1}$$

- The optimization problem in (1) is equivalent to

$$\begin{aligned} &\underset{w,b}{\text{minimizing}} \ \frac{1}{2}\|w\|^2, \\ &\text{subject to } y_i(w^T X_i + b) > 1, i = 1, 2, ..., n. \end{aligned} \tag{2}$$

Hinge Loss of General Linear SVM

How about that the two classes cannot be separated?

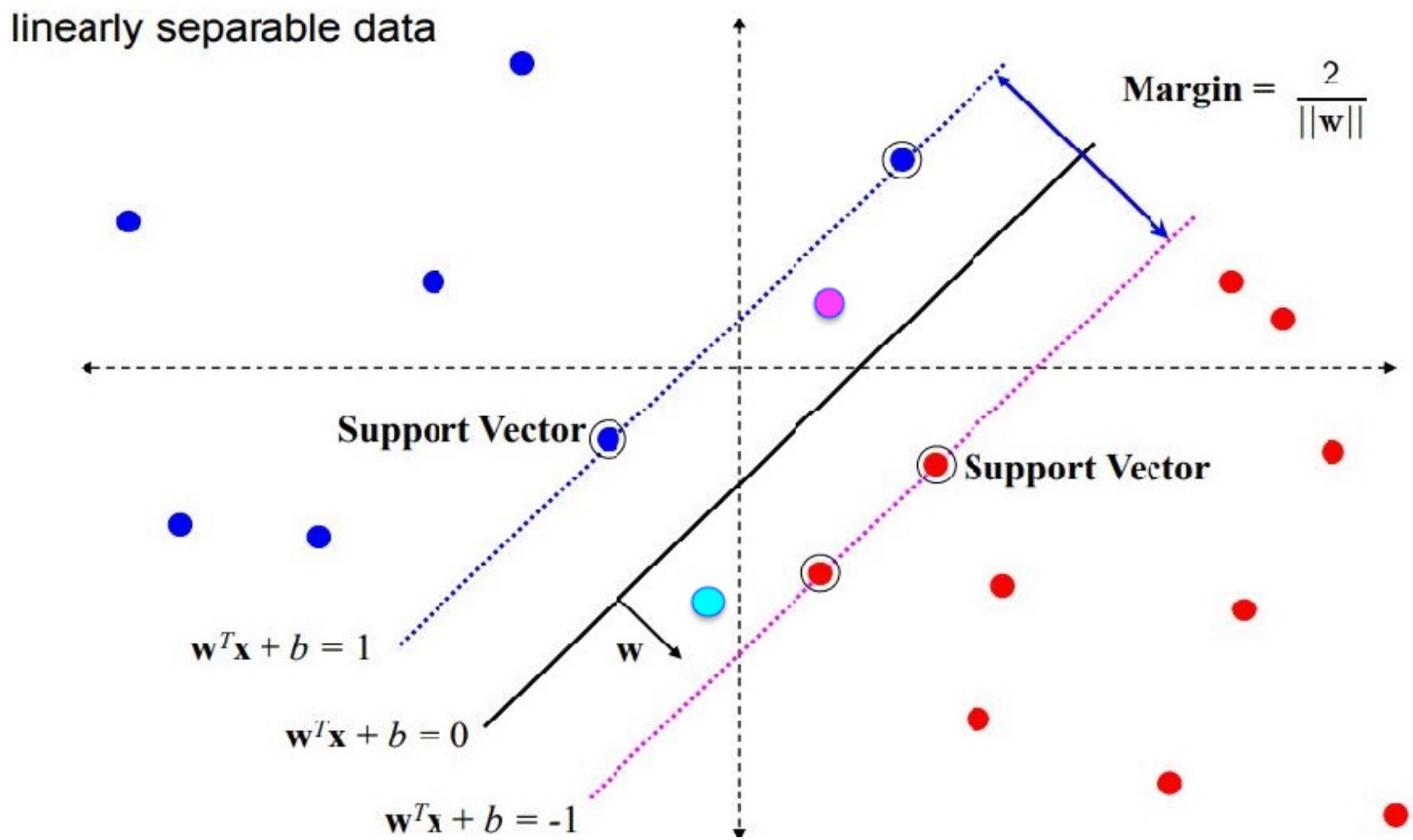


Figure 2: Overview of SVM

Hinge Loss of General Linear SVM (continued)

- $y_i(\mathbf{w}_i^T x + b) \geq 0 \Rightarrow good \Rightarrow Loss = 0$
- $y_i(\mathbf{w}_i^T x + b) \in (0, 1) \Rightarrow not\ good\ enough \Rightarrow Loss = 1 - y_i(\mathbf{w}_i^T x + b)$
- $y_i(\mathbf{w}_i^T x + b) < 0 \Rightarrow bad \Rightarrow Loss = 1 - y_i(\mathbf{w}_i^T x + b)$
- Hinge Loss is defined as

$$L_{\text{Hinge Loss}} = \sum_{i=1}^{n} [1 - y_i(\mathbf{w}_i^T x + b)]_+$$

where, $[z]_+ = max(0, z)$

Subsection 3

Regularization

LASSO Regularization

The least absolute shrinkage and selection operator (LASSO) can set a constraint on the sum of the absolute values of coefficients. By adding the L_1 penalty into the loss function, LASSO conducts a shrinkage process that penalizes the coefficients of features, setting some of them to zero.

- For Logistic Regression, the LASSO estimator is defined as:

$$\underset{\beta_0,\beta}{\operatorname{argmin}}[-\sum_{i=1}^{n}(y_i(\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)})-\log(1+e^{\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)}}))+\lambda\|\beta\|_1],$$

  where $\|\beta\|_1$ is L_1 norm of β, equal to $\sum_{j=1}^{p}|\beta_j|$. The intercept term β_0 is not included in penalizing.

- Similarly, the LASSO estimator of linear SVM is defined as:

$$\underset{b,w}{\operatorname{argmin}}[\sum_{i=1}^{n}[1-y_i(w^T X_i+b)]_+ +\lambda\|w\|_1].$$

Elastic-net Regularization

The elastic-net penalty combines a L_1 penalty with a L_2 penalty: $\lambda[\alpha\|\beta\|_1 + (\frac{1-\alpha}{2})\|\beta\|_2^2]$, where $\|\beta\|_2^2$ is $\sum_{j=1}^{p}\beta_j^2$, called L_2 norm. $\alpha = \frac{\lambda_1}{\lambda_2+\lambda_1}$ and $\lambda = \lambda_2 + \lambda_1$. Then α is the mixing parameter and λ is the tuning parameter. L_1 norm penalty performs feature selection, whereas the L_2 norm penalty allows correlated features to be selected together. Elastic-net estimators of logistic regression and linear SVM are:

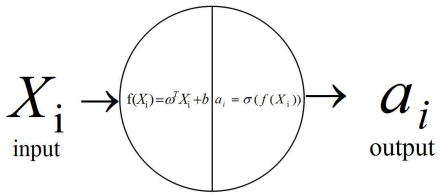
- $\underset{\beta_0,\beta}{\operatorname{argmin}}[-\sum_{i=1}^{n}(y_i(\beta_0 + \sum_{j=1}^{p}\beta_j X_i^{(j)}) - \log(1 + e^{\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)}})) + \lambda(\alpha\|\beta\|_1 + (\frac{1-\alpha}{2})\|\beta\|_2^2]$
- $\underset{b,w}{\operatorname{argmin}}[\sum_{i=1}^{n}[1 - y_i(w^T X_i + b)]_+ + \lambda(\alpha\|w\|_1 + \frac{(1-\alpha)\|w\|_2^2}{2})]$.

Subsection 4

Probabilistic Prediction with SVM's Results

Probabilistic Prediction with SVM's Results

After training SVM models, the outputs consist of relative distance of each observation, $w^T X_i + b$, and the predicted label. SVM model does not generate the probability of each observation getting PD. However, we need probabilistic results to calculate some evaluation metrics. Hence, transforming outputs into probabilistic values is necessary. We use logistic regression, with the relative distance as the input feature, and the phenotype as the response variable.

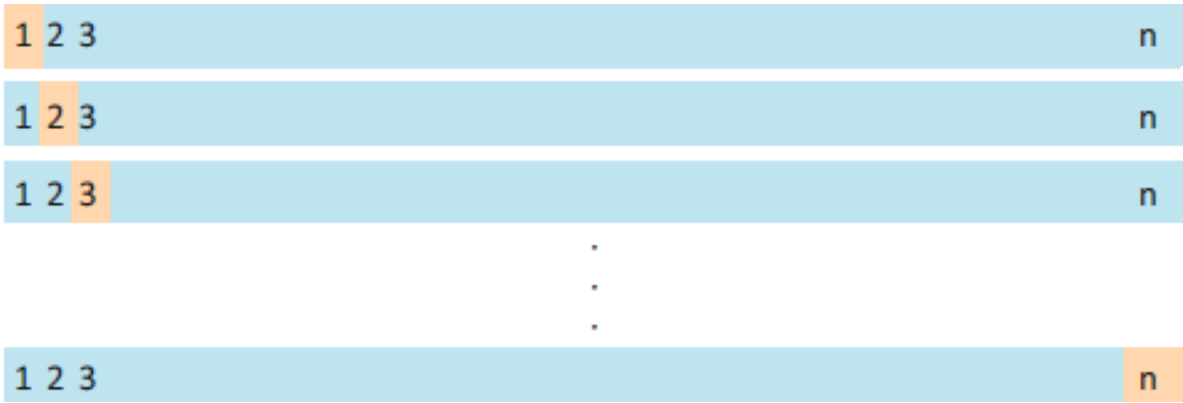


$$a_i = \frac{1}{1 + \exp\left(-(w^T x_i + b)\right)}$$

Section 4

Leave-One-Out Cross-validation

LOOCV



- Keep the 1^{st} sample as test case, use other 326 ones to construct the model. We can obtain coefficients $\beta_1^{[1]}, \beta_2^{[1]}, ..., \beta_{384}^{[1]}, \beta_0^{[1]}$, and the predicted probability of the 1^{st} people, $\hat{p}_{test}^{[1]}$
- Keep the i^{th} sample as the test, use other 326 ones to construct the model. We obtain $\beta_1^{[i]}, \beta_2^{[i]}, ..., \beta_{384}^{[i]}, \beta_0^{[i]}$, and $\hat{p}_{test}^{[i]}$

LOOCV

- After 327 iterations, we get test results of probabilities of all samples: $\hat{p}_{test}^{[1]}, \hat{p}_{test}^{[2]}, ..., \hat{p}_{test}^{[327]}$ and coefficient matrix for 327 iterations:

$$\begin{bmatrix} \beta_1^{[1]} & \beta_2^{[1]} & \dots & \beta_{384}^{[1]} \\ \beta_1^{[2]} & \beta_2^{[2]} & \dots & \beta_{384}^{[2]} \\ \dots & \dots. & \dots & \dots \\ \beta_1^{[327]} & \beta_2^{[327]} & \dots & \beta_{384}^{[327]} \end{bmatrix}$$

- We use $\hat{p}_{test}^{[1]}, \hat{p}_{test}^{[2]}, ..., \hat{p}_{test}^{[327]}$ as the test results of probabilities of all 327 samples, if $\hat{p}_{test}^{[i]} \geq 0.5$, we predict the sample is Parkinson patient, otherwise, the sample is the healthy. Compared with the 327 observed values, we can calculate the evaluation metrics, such as error rates, average minus log probabilities.

Section 5

Evaluation of Predictive Models

Two Evaluation Metrics

- Error rate means the proportion of wrong predictions, which is defined as:

$$ER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i),$$

where y_i is the observed case of sample i, and $\hat{y}_i$ is predicted case of sample i.

- The average of minus log predictive probabilities (AMLP), more commonly called cross-entropy in machine learning textbooks, is:

$$AMLP = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{P}_i(y_i|x_i)).$$

AMLP measures not only the correctness of a point estimate $\hat{y}_i$ but also the level of correctness expressed by $\hat{P}_i(y_i|x_i)$.

Relative Evaluation Metrics

- An example: if there is a dataset with 95% patients and 5% controls, then the baseline error rate and AMLP are 5% and -[0.05log(0.05)+0.95log(0.95)], respectively. If the test error rate of a predictive model is 6%, it looks like very good with only 0.06 error rate. However, even if we use the random guess without any predictors, we can obtain the baseline error rate, 5%. Therefore, the model is not better than the baseline model.
- Baseline prediction is the prediction based on the frequency of observed y_i , without models and predictors. The frequency of $y_i = 1$ is $f_1 = \frac{1}{n}\sum_{i=1}^{n} I(y_i = 1)$, and the frequency of $y_i = 0$ is $f_0 = \frac{1}{n}\sum_{i=1}^{n} I(y_i = 0)$. Then the baseline error rate and AMLP are $ER_{(0)} = min\{f_0, f_1\}$, $AMLP_{(0)} = -[f_0 log(f_0) + f_1 log(f_1)]$, respectively.

Relative Evaluation Metrics (continued)

- In brief, to assess the level of predictivity of a model, we should compare error rate and AMLP with baseline values. Thus, we define the relative error rate as:

$$R^2_{ER} = \frac{ER_{(0)} - ER}{ER_{(0)}}$$

and the relative AMLP as:

$$R^2_{AMLP} = \frac{AMLP_{(0)} - AMLP}{AMLP_{(0)}}$$

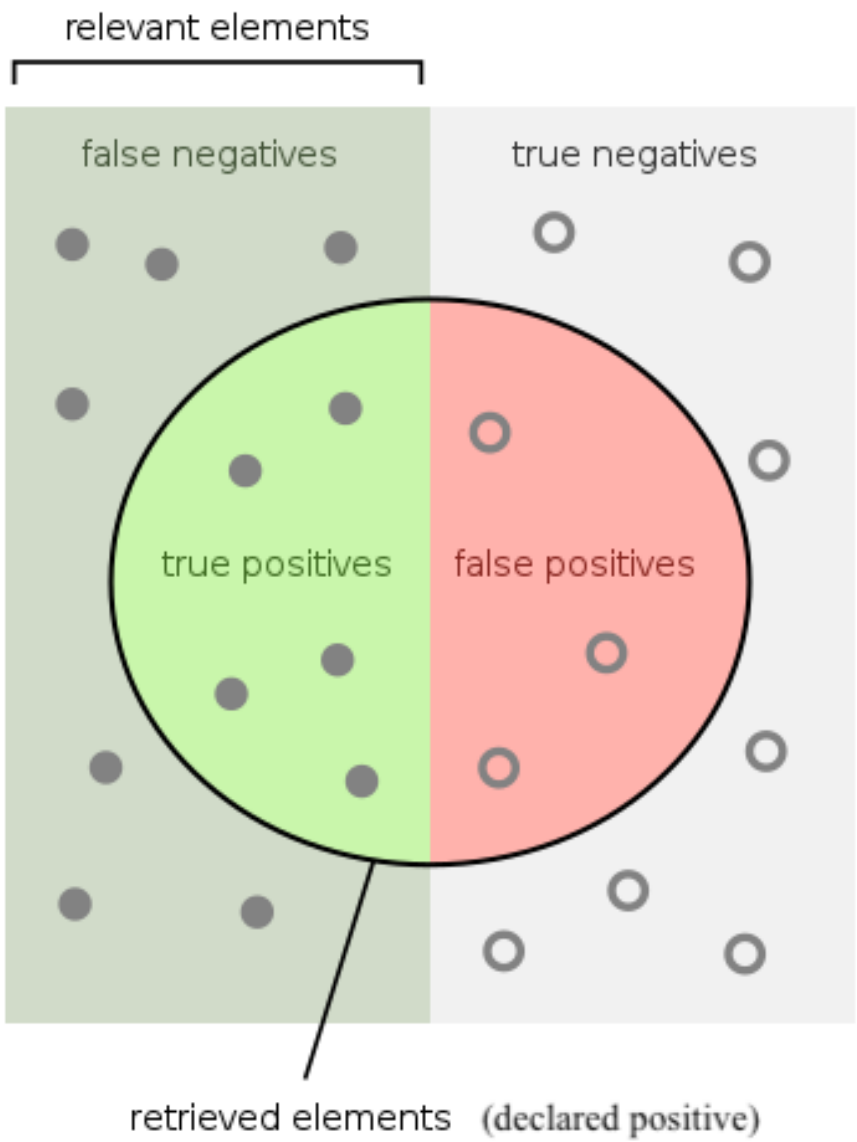
Confusion Matrix for Binary Classification

The predicted labels of all samples are either positive (P) or negative (N). For a sample whose predicted label is P, if its actual label is also P, then it is called a true positive (TP); if the actual label is N, then it is a false positive (FP). Similarly, a true negative (TN) means both the predicted label and the actual label are N, and false negative (FN) means the predicted label is N, whereas its actual label is P. We can summary these notations by building a confusion matrix:

Table 1: Confusion Matrix

		Predicted Labels	
		Class 1	Class0
Actual Labels	Class 1	TP	FN
	Class 0	FP	TN

A Picture for Illustrating Confusion Matrix



AUC

- Sensitivity and Specificity are defined as:

$$Sensitivity = \frac{the\ number\ of\ TP}{the\ number\ of\ TP + the\ number\ of\ FN}$$

$$Specificity = \frac{the\ number\ of\ TN}{the\ number\ of\ TN + the\ number\ of\ FP}$$

- ROC space is defined by One minus Specificity and Sensitivity as x-axis and y-axis, respectively. Suppose c is the threshold, and $\hat{P}_i$ is the predicted probability of a certain sample being positive. If $\hat{P}_i \geq c$, then the predicted label of this sample is P; if $\hat{P}_i < c$ then its predicted label is N. By iteratively using each predicted probability as the threshold, we calculate the corresponding sensitivity and specificity and depict them in the ROC space.
- AUC is the area under the ROC curve. The baseline AUC is 0.5, which can be interpreted as a random guess. A prediction having an AUC closer to 1, is considered superior. The theoretical definition of AUC is the probability that a randomly selected actual positive has a higher test result than a randomly selected actual negative. AUC expresses how much a model is able to distinguish two classes.

AUPR

- Precision and recall are defined as:

$$Precision = \frac{the\ number\ of\ \ TP}{the\ number\ of\ \ TP + the\ number\ of\ \ FP}$$

$$Recall = \frac{the\ number\ of\ \ TP}{the\ number\ of\ \ TP + the\ number\ of\ \ FN}$$

  Recall is another name for sensitivity.
- PRC space is defined by Recall and Precision as x axis and y axis, respectively. Similar with ROC, each point in PRC is positioned by recall as x coordinate and precision as y coordinate. Precision and recall for each point are obtained by using corresponding predicted probability as the threshold. The baseline of PRC is a horizontal line at y=f_1. Here, f_1 is the frequency of positive observations in our dataset. This line divides the precision-recall space into two parts. Curves in the area above the line denote relatively good predictive performance.
- AUPR is the area under the precision-recall curve. The best possible value is 1.0. If a predictive model has a perfect AUPR, it means this model can find all positive samples without incorrectly classifying any negative samples to be positive.

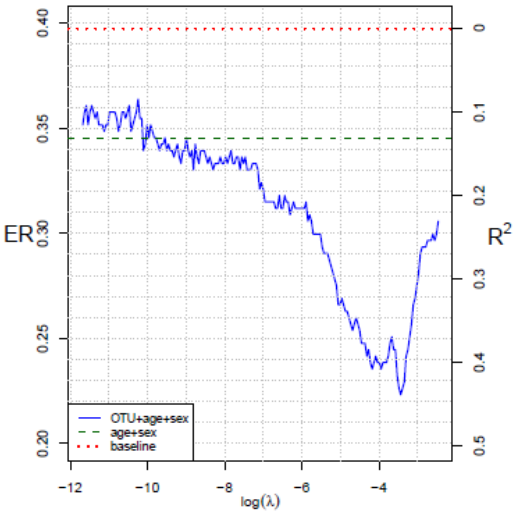
Section 6

Results

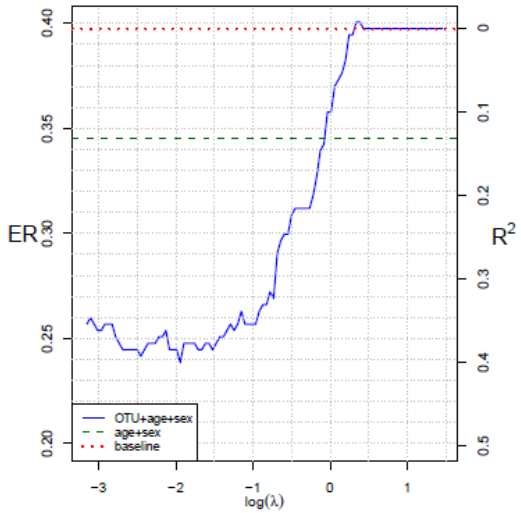
Subsection 1

Predictivity of OTUs for PD

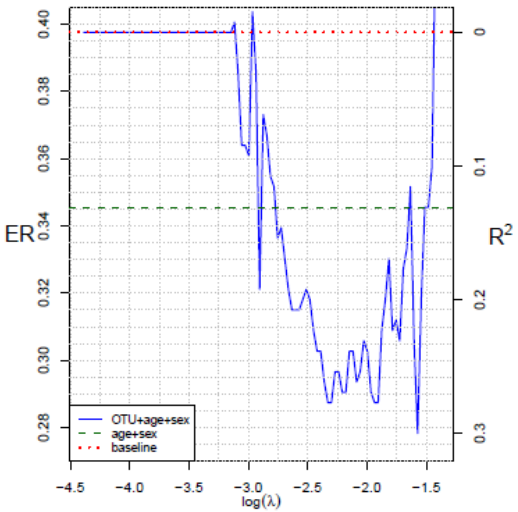
Error Rates of Regularized Logistic Regression and Regularized SVM, as A Function of Different λs



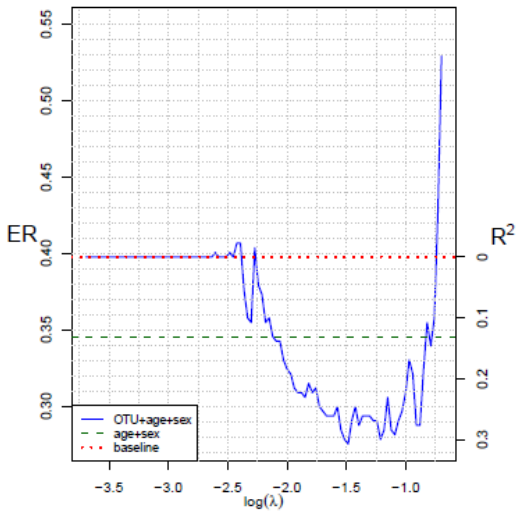
(a) Error Rates of of LR with L_1



(b) Error Rates of of LR with EN



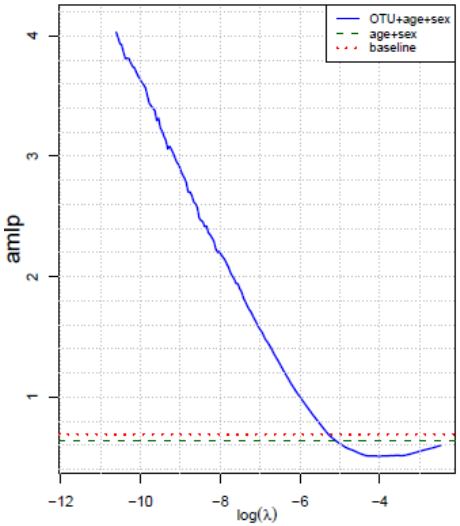
(c) Error Rates of of SVM with L_1



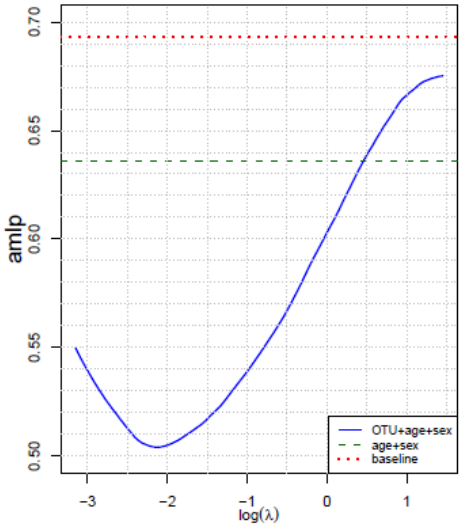
(d) Error Rates of of SVM with EN

- Blue lines show test error rates of regularized logistic regression and SVM, with different values of tuning parameter λ.
- The optimal λ is chosen according to the smallest error rate for each model.
- The smallest error rate of logistic regression with L_1, logistic regression with elastic-net, SVM with L_1 and SVM with elastic-net are 0.223, 0.238, 0.278, 0.275. respectively.

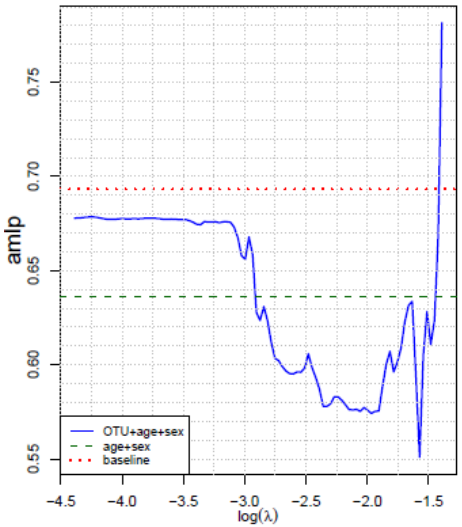
AMLPs of Regularized Logistic Regression and Regularized SVM, as A Function of Different λs



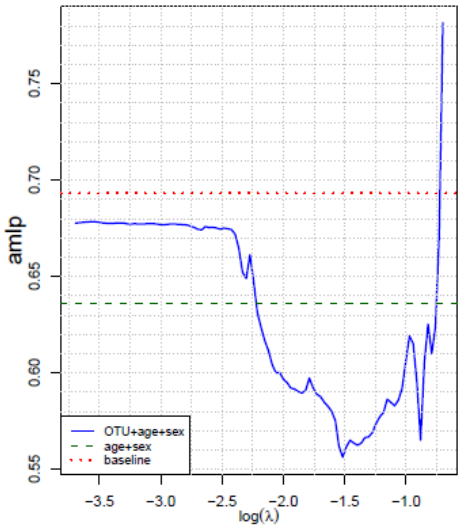
(a) AMLPs of LR with L_1



(b) AMLPs of LR with EN



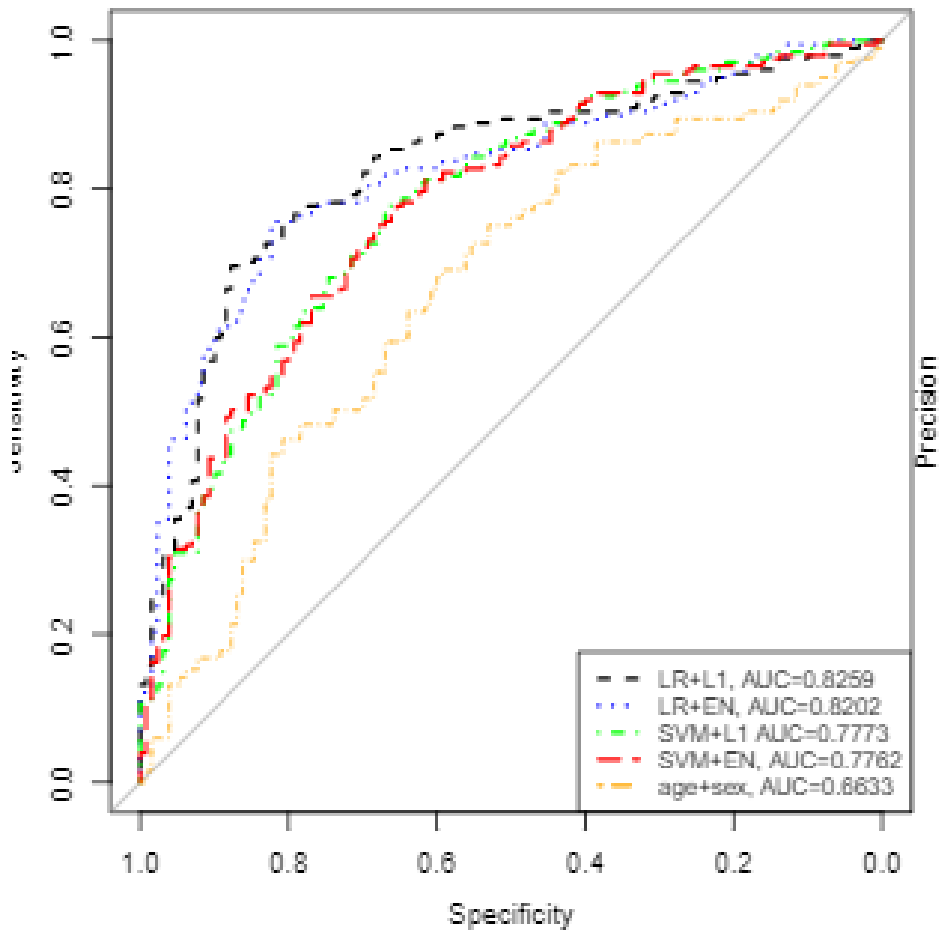
(c) AMLPs of SVM with L_1



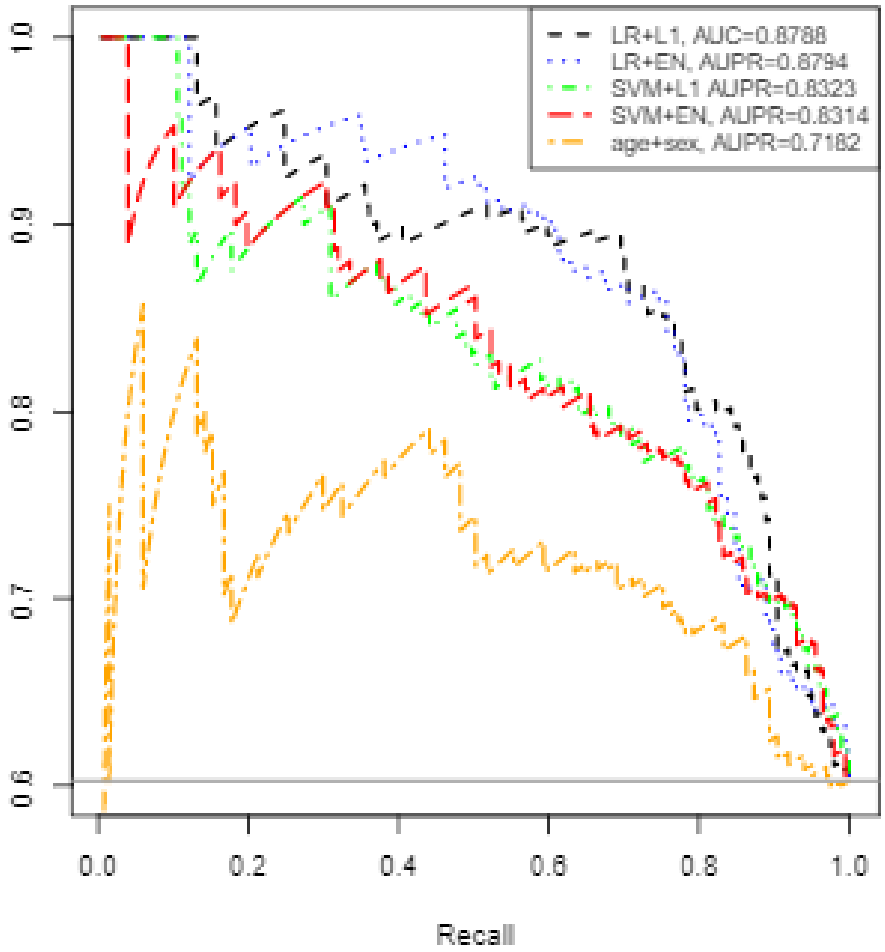
(d) AMLPs of SVM with EN

- Blue lines show AMLP of regularized logistic regression and SVM, with different values of tuning parameter λ.
- Under the optimal λ, the optimal AMLP of logistic regression with L_1, logistic regression with elastic-net, SVM with L_1 and SVM with elastic-net are 0.515, 0.506, 0.550, 0.562, respectively.

ROC and PR Curves of Regularized Logistic Regression and Regularized SVM, with Optimal λ

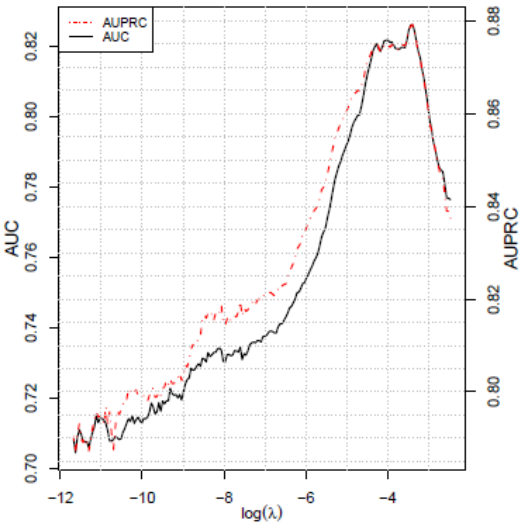


(a) ROC Curves of LR and SVM

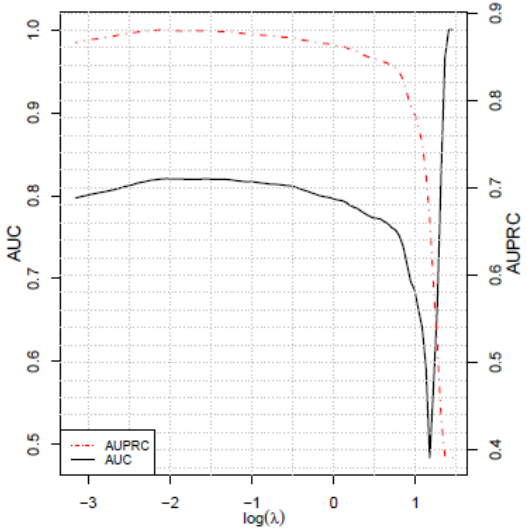


(b) PR Curves of LR and SVM

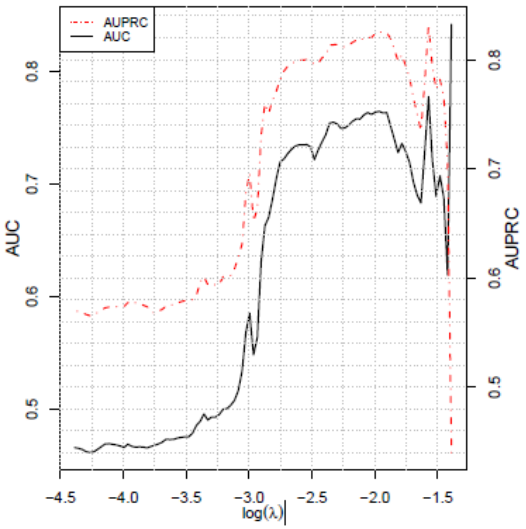
AUC and AUPR of Regularized Logistic Regression and Regularized SVM, as A Function of Different λs



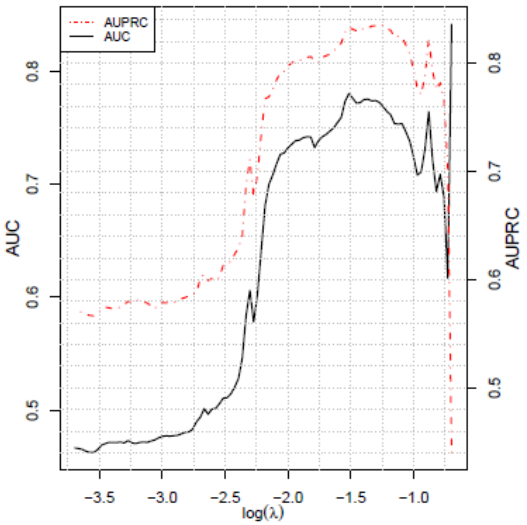
(a) AUC and AUPR of LR with L_1



(b) AUC and AUPR of LR with EN



(c) AUC and AUPR of SVM with L_1



(d) AUC and AUPR of SVM with EN

- Black lines and Red lines show AUC and AUPR, respectively of regularized logistic regression and SVM, with different values of tuning parameter λ.
- Under the optimal λ, the optimal AUC are 0.8259, 0.8202, 0.7773, 0.7762, respectively; the optimal AUPR are 0.8788, 0.8794, 0.8322, 0.8314, respectively.

Summary of Predictive Performance

Table 2: Evaluation of Predictive Models with Respective Optimal λ

ModelMetric	ER (R^2_{ER})	AMLP (R^2_{AMLP})	AUC	AUPR
Baseline (No Predictor)	0.398 (0%)	0.693 (0%)	0.5000	0.6024
LR based on age+sex	0.346 (13.1%)	0.636 (8.20%)	0.6633	0.7182
LR+L_1	**0.223 (43.9%)**	0.515 (25.7%)	**0.8259**	0.8788
LR+EN	0.238 (40.1%)	**0.506 (27.0%)**	0.8202	**0.8794**
SVM+L_1	0.278 (30.1%)	0.550 (20.7%)	0.7773	0.8322
SVM+EN	0.275 (30.8%)	0.562 (18.9%)	0.7762	0.8314

Subsection 2

Selected OTUs

Selection Based on Coefficients

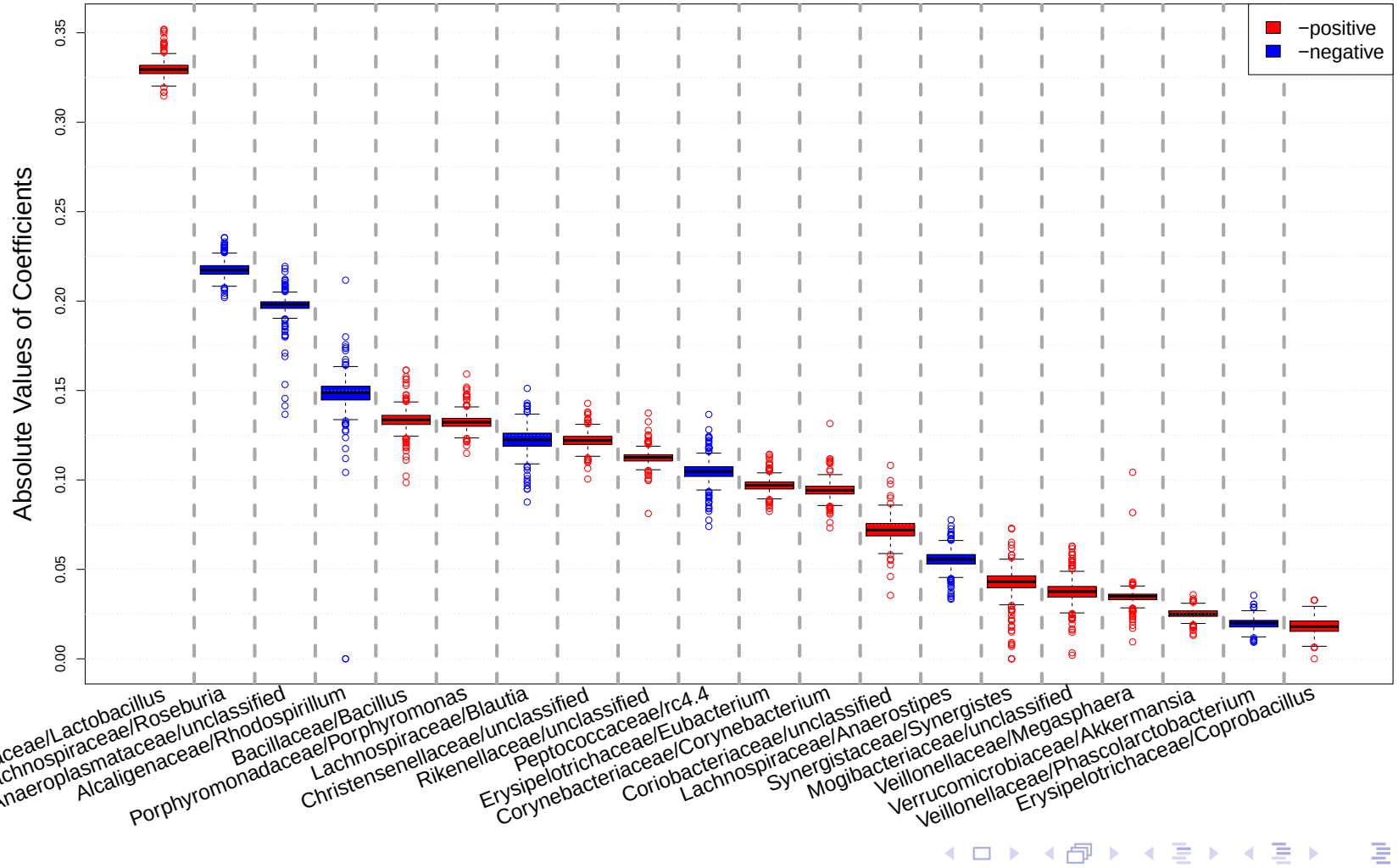
We Choose the best λ based on the smallest error rate. For the best λ for logistic regression with L_1, logistic regression with elastic-net, SVM with L_1, SVM with elastic-net, respectively, we conducted Leave one out cross validation, including 327 iterations because of 327 total samples. Thus we get 327 group of coefficients.

$$\begin{bmatrix} \beta_1^{[1]} & \beta_2^{[1]} & \dots & \beta_{384}^{[1]} \\ \beta_1^{[2]} & \beta_2^{[2]} & \dots & \beta_{384}^{[2]} \\ \dots & \dots. & \dots & \dots \\ \beta_1^{[327]} & \beta_2^{[327]} & \dots & \beta_{384}^{[327]} \end{bmatrix}$$

Each variable has 327 coefficients. For a certain variable, the absolute value of coefficient in a fold may be zero, while in another fold, it might be non-zero. Thus, if the median of absolute values of coefficients from all folds is zero, then the associated variable is eliminated; if the median of absolute values of coefficients is non-zero, the associated variable is retained.

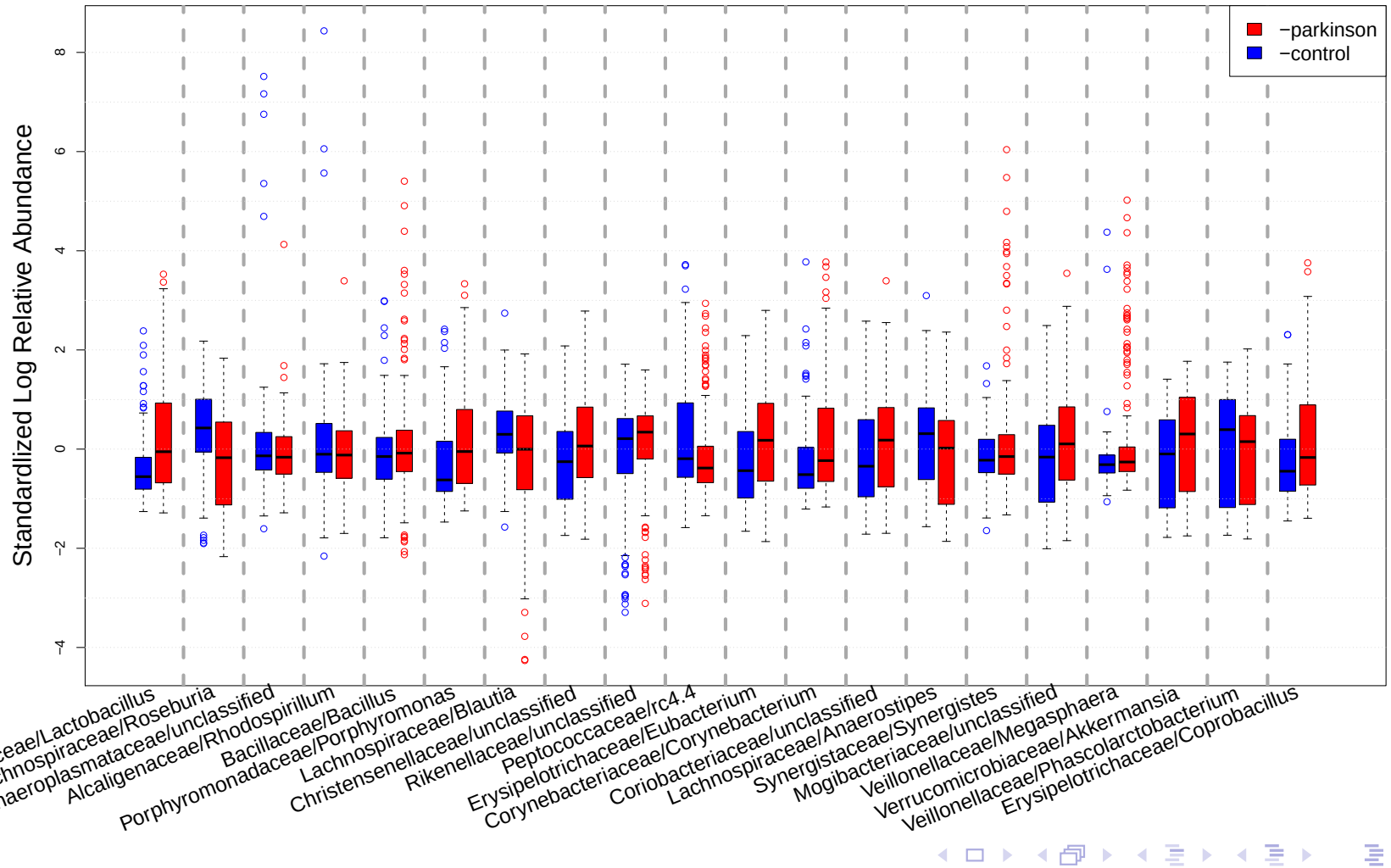
Absolute Coefficients of Selected OTUs by Logistic Regression with L_1

- Logistic Regression with L_1 keeps 25 OTUs.



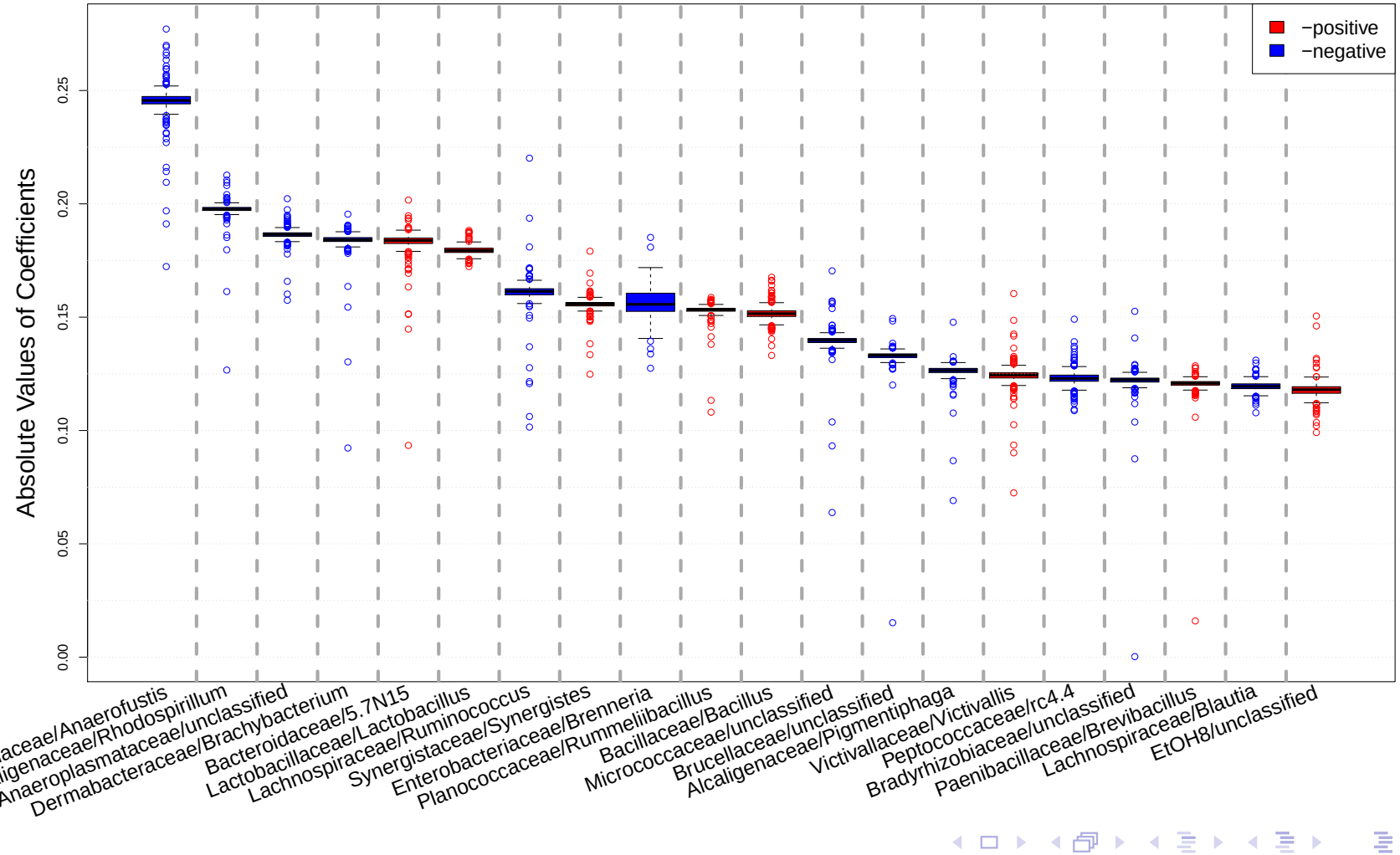
Relative Abundances of Selected OTUs by Logistic Regression with L_1

- Logistic Regression with L_1 keeps 25 OTUs.



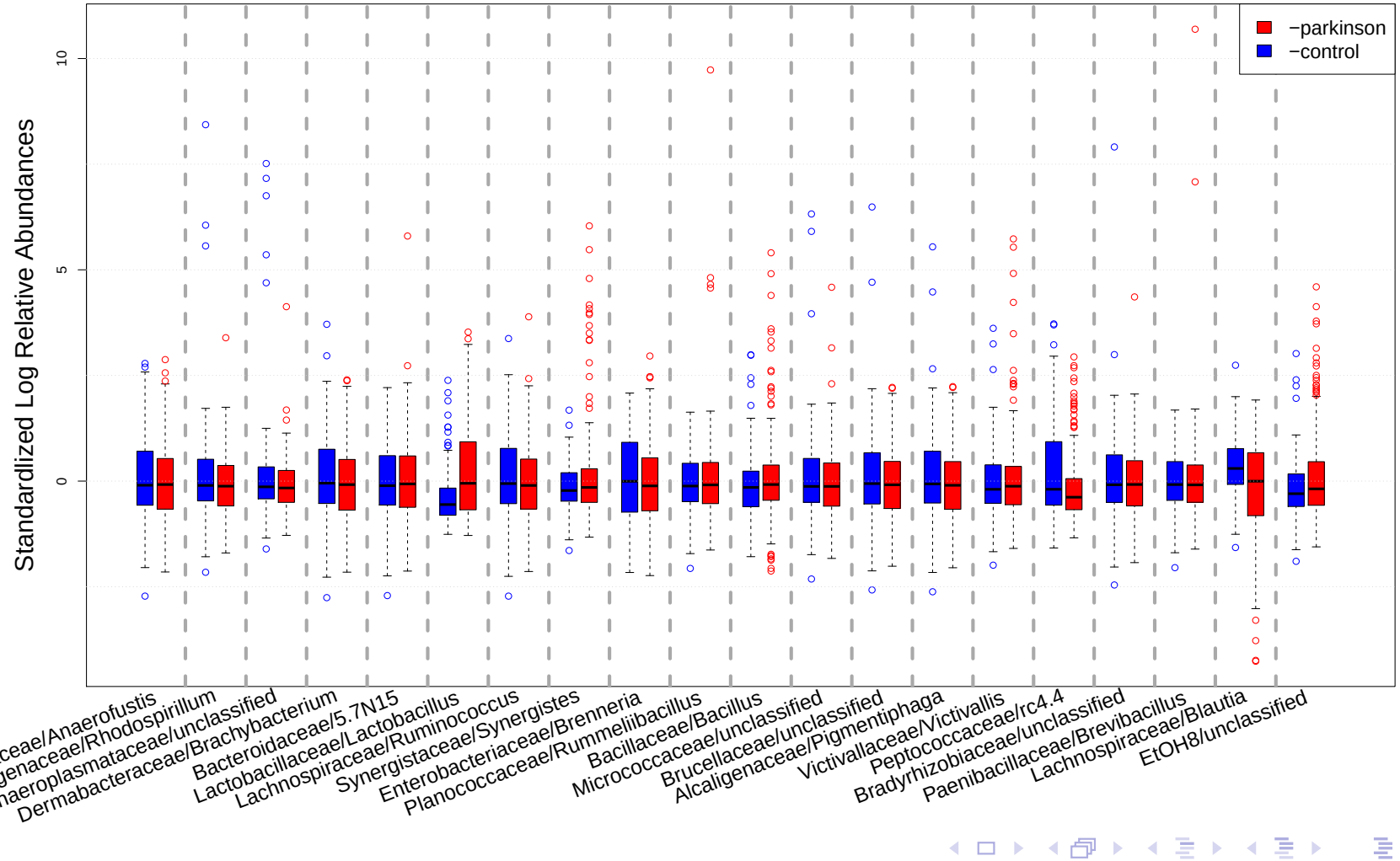
Absolute Coefficients of Selected OTUs by Logistic Regression with Elastic-net

- Logistic Regression with Elastic-net keeps 196 OTUs.



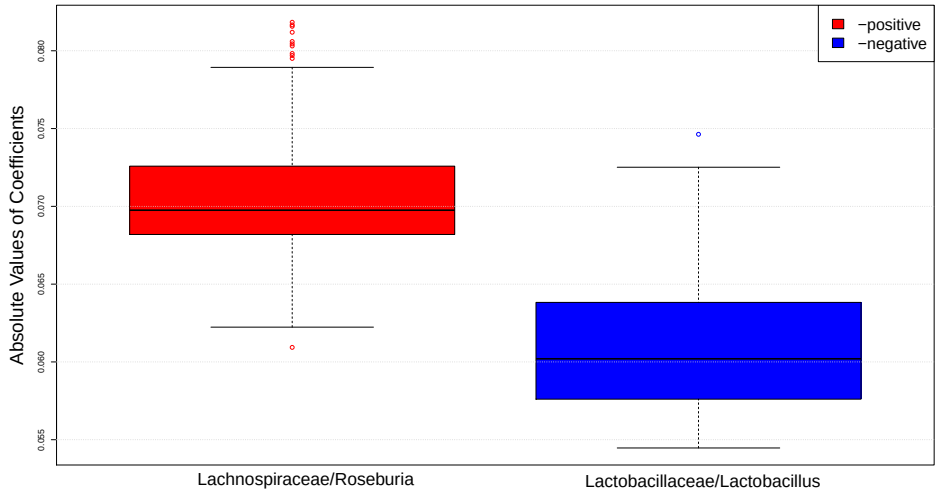
Relative Abundances of Selected OTUs by Logistic Regression with Elastic-net

- Logistic Regression with Elastic-net keeps 196 OTUs.

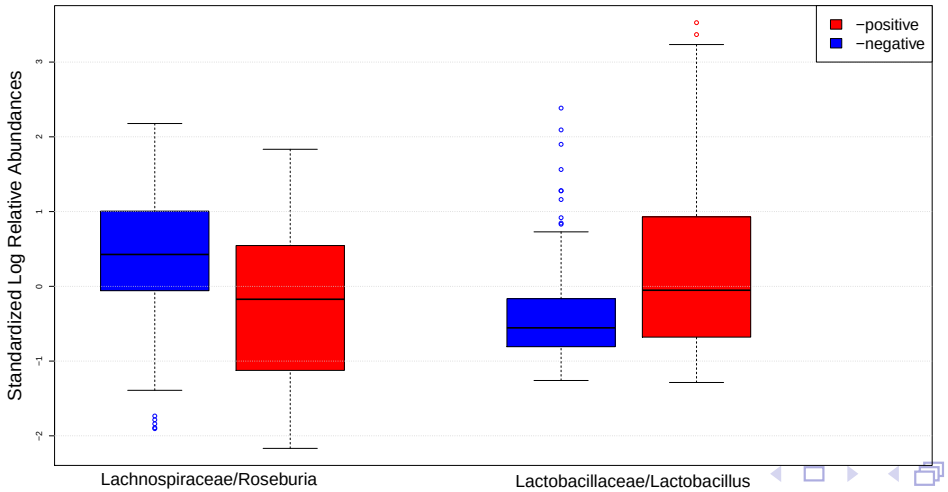


Boxplots of SVM with L_1

- SVM with L_1 keeps 2 OTUs.
- Absolute Coefficients of Selected OTUs by SVM with L_1:

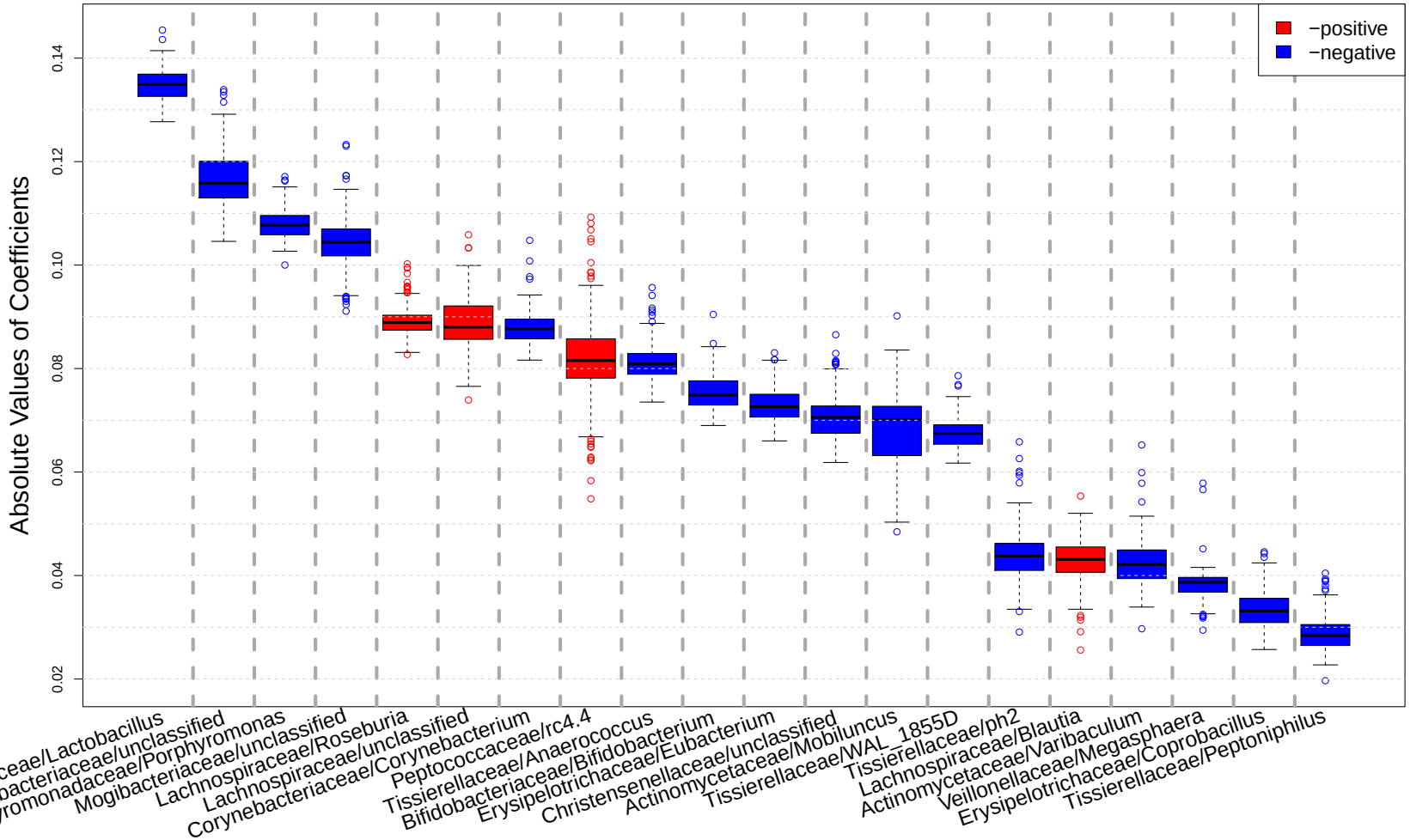


- Relative Abundances of Selected OTUs by SVM with L_1:



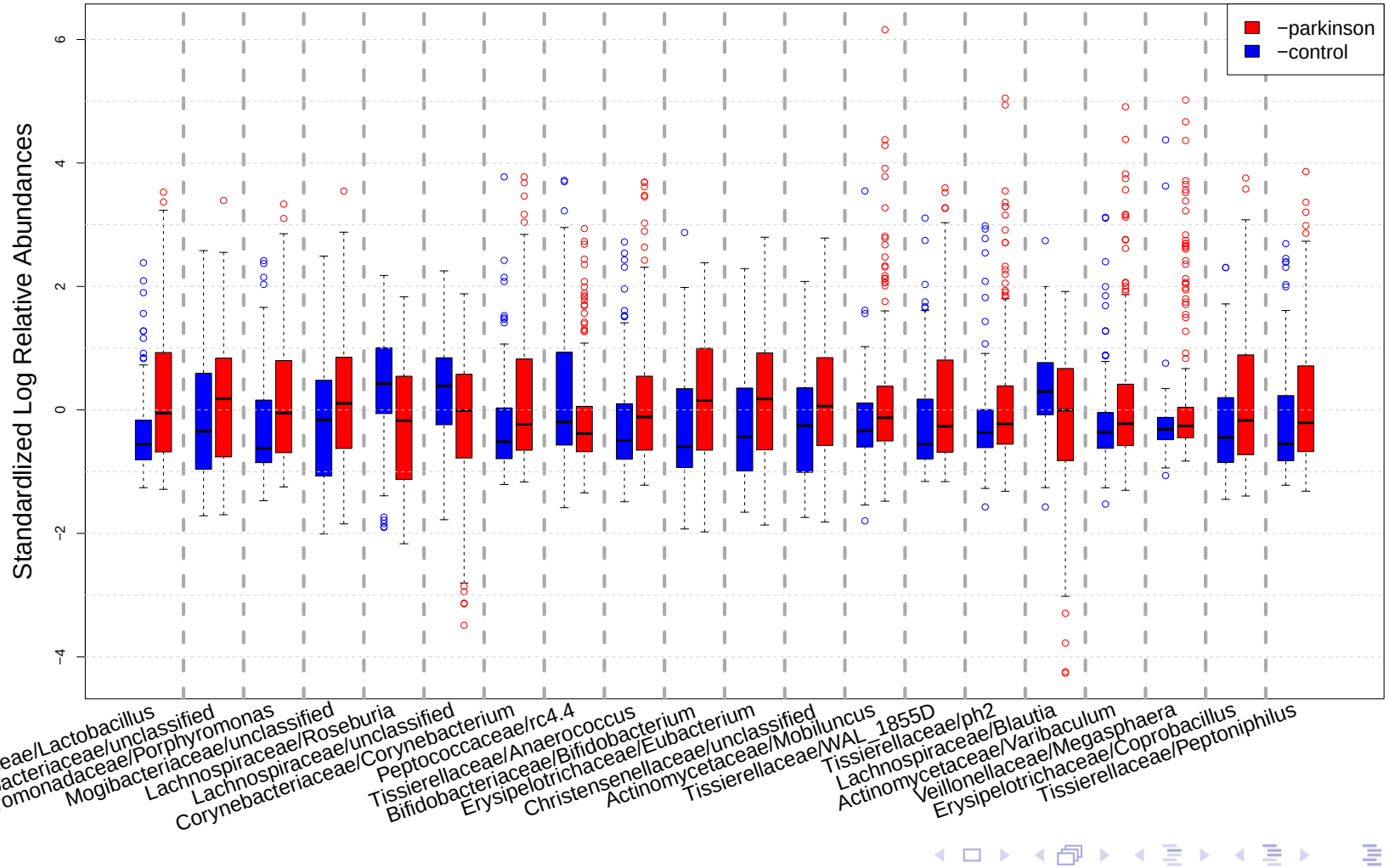
Absolute Coefficients of Selected OTUs by SVM with Elastic-net

- SVM with Elastic-net keeps 27 OTUs.



Relative Abundances of Selected OTUs by SVM with Elastic-net

- SVM with Elastic-net keeps 27 OTUs.



Subsection 3

Comparison of Identified OTUs with Others' Discoveries

Comparison of Selected OTUs with Others' Discoveries

- The most sparse model is SVM with L_1, which retains only 2 OTUs as predictors, **Roseburia** and **Lactobacillus**.
- These two genera are also at the top positions in rankings of selected OTUs by other models. Lactobacillus is ranked first by logistic regression with L_1 penalty. Roseburia is ranked second by logistic regression with L_1.
- The association between the decreased abundance of Roseburia and the presence of PD is also reported by:
  - Janis Bedarf, Falk Hildebrand, et al. Functional implications of microbial and viral gut metagenome change in early-stage l-dopa-naive Parkinson's disease patients. Genome Medicine, 9(1), 2017
  - Ali Keshavarzian, Stefan Green, et al. Colonic bacterial composition in Parkinson's disease. Movement Disorders, 30(10):1351–1360, 2015.
- The association between the increased abundance of Lactobacillus and the presence of PD is also reported by:

Comparison of Selected OTUs with Others' Discoveries (continued)

- Satoru Hasegawa, Sae Goto, et al. Intestinal dysbiosis and lowered serum lipopolysaccharide-binding protein in Parkinson's disease. PLoS One, 10(11):e0142164, 2015.
- Vjacheslav Petrov, Irina Saltykova, et al. Analysis of gut microbiota in patients with Parkinson's disease. Bulletin of Experimental Biology and Medicine volume, 162:734–737, 2017.

Comparison of Selected OTUs with Others' Discoveries (continued)

- **Roseburia** and **Lactobacillus** are both related to the glucose intolerance [1] [2]
- It has been reported that 50% to 80% of patients with Parkinson's disease have glucose intolerance [3].
- Deepseek answers:
  - There is a **well-established association** where Type 2 Diabetes increases the risk of developing Parkinson's Disease by approximately 30-40%.
  - Intriguingly, certain diabetes medications (especially GLP-1 agonists) may reduce PD risk and are being investigated as potential PD treatments.
  - The relationship is complex and bidirectional influences are possible, but the strongest evidence points to T2D as a risk factor for PD.

[1] Yan F, et al. Food Funct. 2019

[2] Nie, K. et al. Front. Cell. Infect. Microbiol. 11, 757718 (2021).

[3] Sandyk R. Int J Neurosci. 1993.

Section 7

Conclusion and Future Work

Conclusion and Future Work

- Our results based on sparse machine learning models provide strong evidence of the connection between PD and the altered gut microbiome. The best AUC is 0.83 and the best AUPR is 0.88. The best reduction in error rate is 44%.
- The genera identified by our predictive models have also been reported in others' studies using other approaches. These genera include Lactobacillus, Blautia, Roseburia, Bifidobacterium, Akkermansia and a genus without a specific name, under the family Christensenellaceae.
- We can repeat our studies with more datasets and consider more variables such as body mass, height, and medications.
- Time-series analysis by using a dataset including data for different PD stages may be a good method, which can bring us a better understanding of the changes in the relative abundance of the gut OTUs during the process of PD.

End

Thank you for your attention!
Questions and comments are welcome!