Variable Selection under Logistic Regression for Compositional Functional Data

Chao Cheng
joint work with Xingdong Feng, Jianhua Hu and Yujie Zhong

School of Statistics and Management
Shanghai University of Finance and Economics

June 12, 2022

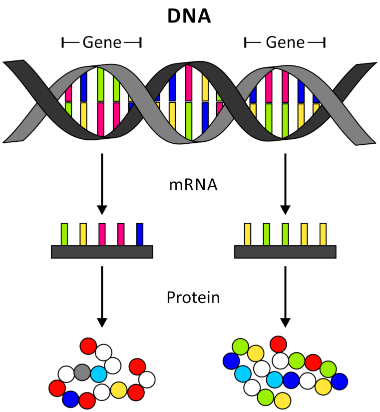
Outline

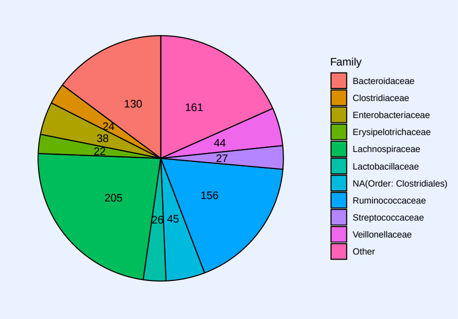
1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

Outline

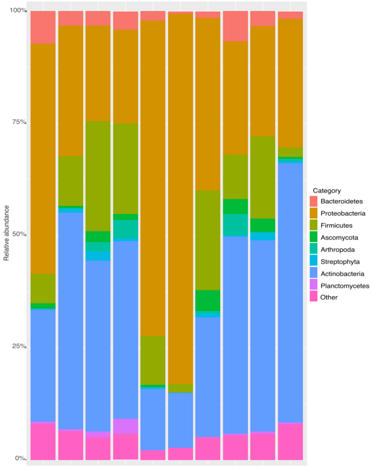
1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

Background


DNA
Gene
Gene
mRNA
Protein


130
161
24
38
22
44
27
205
156
26
45
Family
Bacteroidaceae
Clostridiaceae
Enterobacteriaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
NA(Order: Clostridiales)
Ruminococcaceae
Streptococcaceae
Veillonellaceae
Other

Background



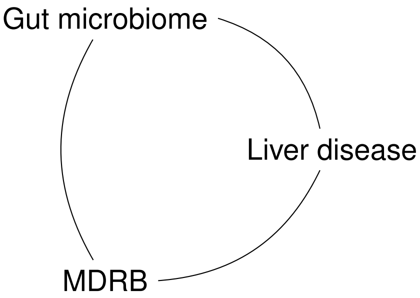
- Sequencing approaches identify numerous microbes.
- The compositional data is more meaningful than raw counts data when studying microbiome.
- Multiple sampling during the study interval naturally results in the formation of functional curves.

Background

Complexity of the data

- High-dimensional data in needs of variable selection
- Complex formation: functional and compositional

Background



- Intestinal microbiome is related to human health in many ways.
- Infections by Multidrug-resistant bacteria (MDRB) remains a leading cause of morbidity and mortality after liver transplantation.

Related work (compositional functional data analysis)

- Log-contrast model: Aitchison (1984); Lin (2014)
- Variable selection for functional data: Ramsay (2002); Fan (2015)
- Compositional functional data analysis via linear regression: Sun (2020)

Outline

1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

Model setup I

Log-contrast model

For compositional data $z_j \in \mathcal{R}_+, j = 1, \cdots, p$ with $\sum_j z_j = 1$, the log-contrast model proposed by (Aitchison , 1984):

$$y = \sum_{j=1}^{p-1} \beta_j \log\left(z_j / z_p\right) + \varepsilon.$$

By introducing $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, the model becomes (Lin , 2014):

$$y = \sum_{j=1}^{p} \beta_j \log z_j + \varepsilon, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0.$$

Model setup II

Logistic model with compositional functional covariates

For binary response $y_i \in \{0, 1\}$, the logistic model for its conditional probability $\pi_i = P\left(y_i = 1 \middle| \boldsymbol{w}_i, X_{ij}(t)\right)$ is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \boldsymbol{w}_i^T \boldsymbol{\delta}_0 + \sum_{j=1}^{p} \int X_{ij}(t)\, \beta_j(t)\, \mathrm{d}t$$

$$\text{s.t.} \sum_{j=1}^{p} \beta_j(t) = 0, \quad \forall t \in [0, 1],$$

where $X_{ij}(t) = \log Z_{ij}(t)$ is the logrithm of the compositional functional data.

Model setup III

Variable selection for functional covariates

$$\sum_{j=1}^{p} P_\lambda\left(\left\|\beta_j\left(\cdot\right)\right\|_2\right),$$

where $\left\|\beta_j\left(\cdot\right)\right\|_2 = \sqrt{\int \beta_j^2\left(t\right)\mathrm{d}t}$ represents L_2-norm of $\beta_j\left(\cdot\right)$.

Model setup IV

Low-rank approximation

Denote $\{b_k(t), k = 1, 2, \cdots\}$ a class of orthornormal basis function on $L^2(\mathrm{d}t)$. Then we have the representation:

$$X_{ij}(t) = \sum_{k=1}^{\infty} \theta_{ijk} b_k(t), \quad \beta_j(t) = \sum_{k=1}^{\infty} \eta_{jk} b_k(t).$$

Apply low-rank approximation by letting $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \cdots, \theta_{ijk_n})^T$ and $\boldsymbol{\eta}_j = (\eta_{j1}, \cdots, \eta_{jk_n})^T$. Then

$$\int X_{ij}(t) \beta_j(t) \,\mathrm{d}t \approx \boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j, \quad \text{and} \quad \|\beta_j(t)\|_2 \approx \|\boldsymbol{\eta}_j\|_2,$$

where $\|\boldsymbol{\eta}_j\|_2 = \sqrt{\boldsymbol{\eta}_j^T \boldsymbol{\eta}_j}$ represents the L_2-norm of a vector.

Model setup V

Low-rank representation of the original model

$$\min S_{\lambda}^{\star}(\beta) = -\frac{1}{n} L(\beta) + \sum_{j=1}^{p} P_{\lambda}\left(\|\eta_j\|_2\right)$$

$$\text{s.t.} \sum_{j=1}^{p} \eta_j = \mathbf{0}_{k_n},$$

where $L(\beta)$ is the MLE of the data, $\beta = \left(\alpha, \delta^T, \eta_1^T, \cdots, \eta_p^T\right)^T$ is the vector of unknown parameters.

Model setup VI

Augmented Lagrangian Multiplier(ALM) method

$$S_\lambda(\beta) = -\frac{1}{n}L(\beta) + \sum_{j=1}^{p} P_\lambda\left(\|\eta_j\|_2\right) + \boldsymbol{\mu}_1^T \sum_{j=1}^{p} \boldsymbol{\eta}_j + \frac{\mu_2}{2}\left\|\sum_{j=1}^{p} \boldsymbol{\eta}_j\right\|_2^2,$$

where $\boldsymbol{\mu}_1$ is the multiplier vector and $\mu_2 > 0$ is the parameter for the augmented term.

Outline

1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

An iterative algorithm

Quadratic approximation at β_0

$$-\frac{1}{n}L(\beta) \approx \frac{1}{n}\left(-L(\beta_0) + (\beta - \beta_0)^T \boldsymbol{X}^T (\pi - \boldsymbol{Y}) + \frac{1}{2}(\beta - \beta_0)^T \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X}(\beta - \beta_0)\right),$$

where $\boldsymbol{W} = \mathrm{diag}(\pi_i(1 - \pi_i)) \leq \frac{1}{4}\boldsymbol{I}_n$.

Quadratic Majorization

Replace $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ with $\boldsymbol{H} = \frac{1}{4}\boldsymbol{X}^T\boldsymbol{X}$.

An iterative algorithm (updating η_{j_0}) I

Apply MM-principle multiple times

$$
\begin{aligned}
S_\lambda\left(\eta_{j_0}\right) \leq\; & \frac{1}{2}\eta_{j_0}^T\left(\frac{d^2_{j_0,max}}{4n}\boldsymbol{I}_{k_n} + \mu_2\boldsymbol{I}_{k_n}\right)\eta_{j_0} \\
& - \eta_{j_0}^T\left(\frac{d^2_{j_0,max}}{4n}\eta_{j_0}^{(m)} - \frac{1}{n}\Theta_{j_0}^T\left(\boldsymbol{\pi}^{(m)} - \boldsymbol{Y}\right) - \boldsymbol{\mu}_1^{(m)} - \mu_2\sum_{j\neq j_0}\eta_j^{(m)}\right) \\
& + P_\lambda\left(\left\|\eta_{j_0}\right\|\right),
\end{aligned}
$$

where $d^2_{j_0,max}$ is the maximum eigen value of $\Theta_{j_0}^T\Theta_{j_0}$ and Θ_{j_0} is the low-rank representation of the j_0th functional covariate.

An iterative algorithm (updating η_{j_0}) II

Local Linear Approximation(LLA)

$$P_\lambda\left(\|\eta_{j_0}\|\right) \approx P_\lambda\left(\left\|\eta_{j_0}^{(m)}\right\|\right) + P'_\lambda\left(\left\|\eta_{j_0}^{(m)}\right\|\right)\left(\|\eta_{j_0}\|_2 - \left\|\eta_{j_0}^{(m)}\right\|_2\right).$$

Updating step for η_{j_0}

$$\eta_{j_0}^{(m+1)} = \frac{1}{d_{j_0,\max}^2/(4n) + \mu_2}\left(1 - \frac{P'_\lambda\left(\left\|\eta_{j_0}^{(m)}\right\|\right)}{\sqrt{\alpha_{j_0}^{(m)T}\alpha_{j_0}^{(m)}}}\right)_+ \alpha_{j_0}^{(m)},$$

where

$$\alpha_{j_0}^{(m)} = \frac{d_{j_0,\max}^2}{4n}\eta_{j_0}^{(m)} - \frac{1}{n}\Theta_{j_0}^T\left(\pi^{(m)} - Y\right) - \mu_1^{(m)} - \mu_2\sum_{j\neq j_0}\eta_j^{(m)}.$$

Implementations in R package I

Sample weighting

$$L_v = \sum_{i=1}^{n} v_i \left(y_i \pi_i + (1 - y_i)(1 - \pi_i)\right),$$

where $v_i, i = 1, \cdots, n$ is a set of positive weights.

Implementations in R package II

Flexibility and robustness

within-group standardization

multiplier vector

$$S_n = \frac{-1}{a} L_v + \sum_{j=1}^{p} c_j P_{\lambda_n}\left(b_j \|\boldsymbol{\eta}\|_2\right) + \boldsymbol{\mu}_1^T \sum_{j=1}^{p} \boldsymbol{\eta}_j + \frac{\mu_2}{2} \left\| \sum_{j=1}^{p} \boldsymbol{\eta}_j \right\|_2^2 + \sum_{j=1}^{h+pk_n} \frac{r_j}{2} \beta_j^2,$$

weights for different modules

robust term

Implementations in R package III

Within-group orthonormalization

$$S_n = \frac{-1}{a} L_v + \sum_{j=1}^{P} c_j P_{\lambda_n}\left(b_j \left\|\tilde{\eta}_j\right\|_2\right) + \boldsymbol{\mu}_1^T \sum_{j=1}^{P} \boldsymbol{T}_j \tilde{\boldsymbol{\eta}}_j + \frac{\mu_2}{2} \left\|\boldsymbol{T}_j \tilde{\boldsymbol{\eta}}_j\right\|_2^2 + \sum_{j=1}^{h+\sum_k m_k} \frac{r_j}{2} \beta_j^2,$$

where $\tilde{\eta}_j$ is the parameters after within-group orthonormalization. $\boldsymbol{T}_j = \sqrt{a} \boldsymbol{V}_j \boldsymbol{D}_j^{-1}$ is the transforming matrix satisfying $\eta_j = T_j \tilde{\eta}_j$. Covariates after transforming $\tilde{\boldsymbol{\Theta}}_j = \boldsymbol{\Theta}_j \boldsymbol{T}_j$ satisfies $\frac{1}{a} \tilde{\boldsymbol{\Theta}}_j^T \tilde{\boldsymbol{\Theta}}_j = \boldsymbol{I}$.

Implementations in R package IV

FLiRTI(James , 2009) procedure

Pursuit the functional parameters in simple form of curves:

$$S_n = \text{Logistic} + \sum_{j=1}^{q}\sum_{k=1}^{T} c_{jk} P_{\lambda}\left(b_{jk}\gamma_{(1),jk}\right) + \boldsymbol{\mu}_1^T \sum_{j=1}^{q}\gamma_{(1),j} + \frac{\mu_2}{2}\left\|\sum_{j=1}^{q}\gamma_{(1),j}\right\|_2^2 .$$

Outline

1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

Convergence of the algorithm

Checking KKT conditions we know that

Theorem

The objective function coupled with Lasso penalty strictly descents during the iteration. If the current solution $\alpha^{(m+1)}$, $\delta^{(m+1)}$, $\eta_j^{(m+1)}, j = 1, \cdots, p$ and $\mu_1^{(m+1)}$ stay unchanged after one complete iteration, then the algorithm reaches the minimum point of the objective function.

Consistency of the Lasso estimator I

Theorem (Consistency of the Lasso estimator)

If Assumption 1–3 hold and $pk_n = o\left(e^n\right)$, $\lambda \to 0$. Also there exists constant c such that $\lambda > c\sqrt{k_n \log\left(pk_n\right)/n}$. Then with probability approaching to 1, the Lasso estimator $\hat{\xi}^{glasso}$ satisfies

$$\left\|\hat{\xi}^{glasso} - \xi^\star\right\|_2 \leq 2\left(1+\phi\right)\left(2+\phi\right)\frac{\sqrt{q_n}\lambda}{\kappa},$$

as $n \to \infty$, where ϕ is any positive constant and $\xi^\star$ is the real underlying parameters.

Consistency of the Lasso estimator II

Corollary

Under the condition that the previous theorem holds, and additionally the minimal signal strength Assumption 4 holds, the with probability approaching to 1,

$$\min_{j\in\{1,\cdots,q_n\}} \left\|\hat{\eta}_j^{glasso}\right\|_2 \geq 2\left(1+\phi\right)\left(2+\phi\right)\frac{\sqrt{q_n}\lambda}{\kappa},$$

and

$$\max_{j\in\{q_n+1,\cdots,p\}} \left\|\hat{\eta}_j^{glasso}\right\|_2 \leq 2\left(1+\phi\right)\left(2+\phi\right)\frac{\sqrt{q_n}\lambda}{\kappa}.$$

as $n \to \infty$.

Necessary assumptions I

1. There uniformly exist positive constants K and R, independent of n, such that for $1 \leq i \leq n$, $1 \leq j \leq p$, $1 \leq k \leq k_n$ and $m = 2, 3, \cdots,$

$$\mathrm{E}\left|\theta_{ij,k}\left(y_i - \pi_i\left(\xi^\star, \boldsymbol{\Theta}\right)\right)\right|^m \leq (m!/2)\, K^{m-2} R^2,$$

   where we use $\pi_i\left(\xi^\star, \boldsymbol{\Theta}\right)$ to emphasize that the computation of π_i relies on underlying parameters $\xi^\star$ and data $\boldsymbol{\Theta}$.

2. There exists a positive constant κ such that

$$\inf\left\{\frac{\sqrt{\Delta^T \boldsymbol{\Theta}^T \boldsymbol{W}\left(\xi^\star, \boldsymbol{\Theta}\right) \boldsymbol{\Theta}\Delta}}{\sqrt{n}\,\|\Delta\|_2} : \boldsymbol{C}^T\Delta = \mathbf{0}, \sum_{j=q_n+1}^{p} \|\Delta_j\|_2 \leq (1+\phi) \sum_{j=1}^{q_n} \|\Delta_j\|_2\right\} = \kappa > 0,$$

   where $\boldsymbol{\Theta} = \left(\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_p\right)$ is the $n \times (pk_n)$ low-rank representation of p functional covariates.

Necessary assumptions II

3. $\mathrm{E}q_3(\xi, \theta)\|\theta\|_{2,\infty}^3$ is bounded in the neighbourhood centered at $\xi^\star$, where $q_3(\xi, \theta)$ represents the 3rd order derivative of the link function of logistic regression.

4. The real underlying model parameters satisfy

$$\min_{j\in\{1,\cdots,q_n\}} \left\|\eta_j^\star\right\|_2 \geq 4(1+\phi)(2+\phi)\sqrt{q_n}\lambda.$$

Oracle property when applying non-convex penalty

Theorem

If Assumption 1–3, 5–7 hold, and $\lambda = o\left(n^{-(1-c_2)/2}\right)$, $q_n k_n \sqrt{k_n/n} = o(\lambda)$, $\sqrt{k_n}\log p = o(n\lambda)$ *and* $n\lambda/\sqrt{k_n} \to \infty$ *are satisfied, then there exists a local minimum* $\left(\hat{\eta}_1^T, \cdots, \hat{\eta}_p^T\right)^T$ *of the objective function coupled with SCAD or MCP penalty such that*

$$P\left(\left(\hat{\eta}_1^T, \cdots, \hat{\eta}_p^T\right)^T = \left(\hat{\eta}_1^{or\ T}, \cdots, \hat{\eta}_p^{or\ T}\right)^T\right) \to 1$$

as $n \to \infty$, *where* $\left(\hat{\eta}_1^{or\ T}, \cdots, \hat{\eta}_p^{or\ T}\right)^T$ *denotes the oracle estimator.*

Necessary assumptions

5. There exist two positive constants C_1 and C_2 uniformly for $j \in \{1, \cdots, q_n\}$, such that

$$0 < C_1 \leq \lambda_{min}\left(\frac{1}{n}\Theta_j^T \boldsymbol{W}\left(\xi^\star, \Theta\right)\Theta_j\right) \leq \lambda_{max}\left(\frac{1}{n}\Theta_j^T \boldsymbol{W}\left(\xi^\star, \Theta\right)\Theta_j\right) \leq C_2,$$

   where $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ represents the minimum and maximum eigen value of a given matrix. Also it's assumed that $\max_{1\leq i\leq n}\left\|\left(\boldsymbol{\theta}_{i1}^T, \cdots, \boldsymbol{\theta}_{iq_n}^T\right)\right\|_2 = O_p\left(\sqrt{q_n k_n}\right)$ and there exists a constant M_1 such that $\max_{j,k}\mathrm{E}\left|\theta_{\cdot jk}\right| \leq M_1$ for $j \in \{q_n + 1, \cdots, p\}$, $k \in \{1, \cdots, k_n\}$.
6. There exists a positive constant c_1 such that $0 \leq c_1 < \frac{1}{3}$ and $q_n k_n = O\left(n^{c_1}\right)$.
7. There exists a positive constant c_2 such that $2c_1 < c_2 < 1$ and

$$n^{(1-c_2)/2}\min_{1\leq j\leq q_n}\left\|\boldsymbol{\eta}_j^\star\right\|_2 \geq M_2.$$

Outline

1 Introduction

2 Model setup

3 An iterative algorithm

4 Theoretical properties

5 Simulations

6 Real world application

Simulation settings I

Data generation

$$\bar{W}_{ij}(t) = \exp\left(\boldsymbol{w}_{ij}^{T}\boldsymbol{c}(t) + \varepsilon_{ij}(t)\right), \quad i = 1, \cdots, n, \quad j = 1, \cdots, p,$$

where $\boldsymbol{c}(t)$ is a set of Fourier basis, $\varepsilon_{ij}(t)$ follows $N(0, 0.5^2)$ at each time points. The functional counting data is acquired by $W_{ij}(t) = \lfloor \bar{W}_{ij}(t) \rfloor$. Hence

$$X_{ij}(t) = \log\left(\frac{W_{ij}(t) + 1}{\sum_{j=1}^{p} W_{ij}(t) + p}\right), \quad i = 1, \cdots, n, \quad j = 1, \cdots, p.$$

Simulation settings II

Low-rank approximation

- First q_n terms of $\beta_j(t)$ are generated from $\beta_j = \boldsymbol{\eta}_j^T \boldsymbol{c}(t)$, where $\boldsymbol{\eta}_j$ is centerd so that $\sum_{j=1}^{q_n} \boldsymbol{\eta}_j = \mathbf{0}$. The remaining $\beta_j(t), j = q_n + 1, \cdots, p$ are set to constant 0.
- Orthonormalized B-splineRedd (2011) is applied for low-rank approximation.

Magnitude of the simulation

- Sample size $n \in \{75, 100, 150, 250\}$, number of functional covariates $p \in \{50, 500\}$.
- Number of important covariates $q_n \in \{8, 10\}$, number of basis functions $k_n = 6$.

Simulation settings III

Candidate methods

- Oracle model
- Proposed model, coupled with MCP and Lasso penalty
- GGL (Generalized linear regression with Group Lasso, Yang (2014))
- ZINB (Zero Inflated Negative Binomial model, Zhang (2018))

Simulation results (low-dimensional settings) I

Table 1: $n = 150$, $p = 50$, $q = 10$

Model	Criteria	AUC	MSE_δ	$\lvert\sum\hat{\eta}\rvert_\infty$	FP	FN	FDR
Oracle	-	0.947(0.039)	0.711(0.439)	$< 10^{-5}$	0(0)	0(0)	0(0)
MCP	BIC	0.881(0.081)	0.847(0.477)	$< 10^{-5}$	1.010(0.969)	3.860(1.059)	0.138(0.131)
MCP	CV	0.827(0.095)	0.794(0.272)	$< 10^{-5}$	0.150(0.411)	6.420(0.987)	0.035(0.093)
Lasso	BIC	0.870(0.070)	0.803(0.456)	$< 10^{-5}$	1.410(1.156)	3.840(0.801)	0.172(0.130)
Lasso	CV	0.843(0.075)	0.766(0.250)	$< 10^{-5}$	18.030(3.465)	0.790(0.715)	0.657(0.046)
GGL	CV	0.877(0.069)	2.046(1.977)	0.546	12.510(6.522)	1.380(1.277)	0.538(0.172)
ZINB	fdr0.05	0.599(0.092)	1.208(1.255)	0.107	0(0)	9.810(0.465)	0(0)

Simulation results (low-dimensional settings) II

Table 2: $n = 250$, $p = 50$, $q = 10$

Model	Criteria	AUC	MSE_δ	$\lvert\sum\hat{\eta}\rvert_\infty$	FP	FN	FDR
Oracle	-	0.973(0.020)	0.411(0.262)	$< 10^{-5}$	0(0)	0(0)	0(0)
MCP	BIC	0.954(0.030)	0.483(0.204)	$< 10^{-5}$	0.350(0.672)	1.950(0.770)	0.040(0.076)
MCP	CV	0.927(0.052)	0.587(0.225)	$< 10^{-5}$	0.150(0.557)	3.670(1.596)	0.016(0.054)
Lasso	BIC	0.949(0.032)	0.503(0.194)	$< 10^{-5}$	0.700(0.847)	2.230(0.777)	0.074(0.086)
Lasso	CV	0.884(0.053)	0.617(0.167)	$< 10^{-5}$	23.750(3.020)	0.220(0.416)	0.706(0.031)
GGL	CV	0.934(0.037)	1.505(1.330)	0.472	14.850(7.124)	0.590(0.698)	0.571(0.145)
ZINB	fdr0.05	0.573(0.076)	1.111(0.943)	0.078	0.010(0.010)	9.880(0.327)	0.010(0.100)

Simulation results (high-dimensional settings) I

Table 3: $n = 150$, $p = 500$, $q = 8$

Model	Criteria	AUC	MSE_δ	$\lvert\sum\hat{\eta}\rvert_\infty$	FP	FN	FDR
Oracle	-	0.947(0.031)	1.093(0.704)	$< 10^{-5}$	0(0)	0(0)	0(0)
MCP	BIC	0.803(0.100)	0.769(0.516)	$< 10^{-5}$	2.920(1.839)	3.200(1.325)	0.362(0.209)
MCP	CV	0.738(0.098)	0.518(0.214)	$< 10^{-5}$	0.580(0.673)	5.280(1.230)	0.150(0.168)
Lasso	BIC	0.804(0.091)	0.587(0.453)	$< 10^{-5}$	3.615(1.670)	2.769(1.231)	0.398(0.157)
Lasso	CV	0.760(0.061)	0.466(0.175)	$< 10^{-5}$	31.173(8.740)	0.981(1.180)	0.804(0.059)
GGL	CV	0.794(0.093)	3.727(4.126)	0.440	26.580(16.519)	1.460(1.854)	0.704(0.225)
ZINB	fdr0.05	0.571(0.061)	0.925(1.330)	0.033	4.260(2.039)	7.900(0.364)	0.966(0.149)

Simulation results (high-dimensional settings) II

Table 4: $n = 250$, $p = 500$, $q = 8$

Model	Criteria	AUC	MSE_δ	$\lvert\sum\hat{\eta}\rvert_\infty$	FP	FN	FDR
Oracle	-	0.954(0.023)	1.892(0.947)	$< 10^{-5}$	0(0)	0(0)	0(0)
MCP	BIC	0.951(0.047)	0.233(0.211)	$< 10^{-5}$	0.750(1.065)	0.062(0.250)	0.077(0.101)
MCP	CV	0.929(0.057)	0.185(0.127)	$< 10^{-5}$	0.312(0.602)	0.938(1.482)	0.033(0.063)
Lasso	BIC	0.937(0.036)	0.183(0.127)	$< 10^{-5}$	2.111(2.147)	0.444(0.726)	0.194(0.150)
Lasso	CV	0.840(0.068)	0.314(0.111)	$< 10^{-5}$	39.444(2.963)	0.000(0.000)	0.831(0.011)
GGL	CV	0.921(0.045)	1.442(1.232)	0.322	28.000(19.280)	0.062(0.250)	0.707(0.185)
ZINB	fdr0.05	0.590(0.076)	0.605(0.127)	0.023	5.438(2.449)	7.938(0.250)	0.992(0.031)

Simulation results (settings that mimic real data)

Table 5: $n = 150$, $p = 500$, $q = 8$. Subjects are heterogeneous, mimicing the pattern in real data application.

Model	Criteria	AUC	MSE$_\delta$	$\lvert\sum \hat{\eta}\rvert_\infty$	FP	FN	FDR
Oracle	-	0.883(0.055)	2.080(2.588)	$< 10^{-5}$	0(0)	0(0)	0(0)
MCP	BIC	0.797(0.081)	0.630(0.837)	$< 10^{-5}$	5.000(1.732)	5.348(0.775)	0.637(0.135)
MCP	CV	0.797(0.085)	0.782(0.551)	$< 10^{-5}$	1.565(2.063)	6.391(1.438)	0.257(0.283)
Lasso	BIC	0.908(0.026)	0.287(0.130)	$< 10^{-5}$	2.000(1.000)	5.200(0.837)	0.410(0.175)
Lasso	CV	0.780(0.047)	1.664(0.742)	$< 10^{-5}$	20.400(9.764)	3.800(1.304)	0.811(0.063)
GGL	CV	0.815(0.070)	1.604(1.306)	0.758	18.913(16.673)	4.565(1.343)	0.737(0.216)
ZINB	fdr0.05	0.773(0.082)	$> 4 \times 10^3$	1.326	0.522(0.790)	6.217(1.043)	0.159(0.259)

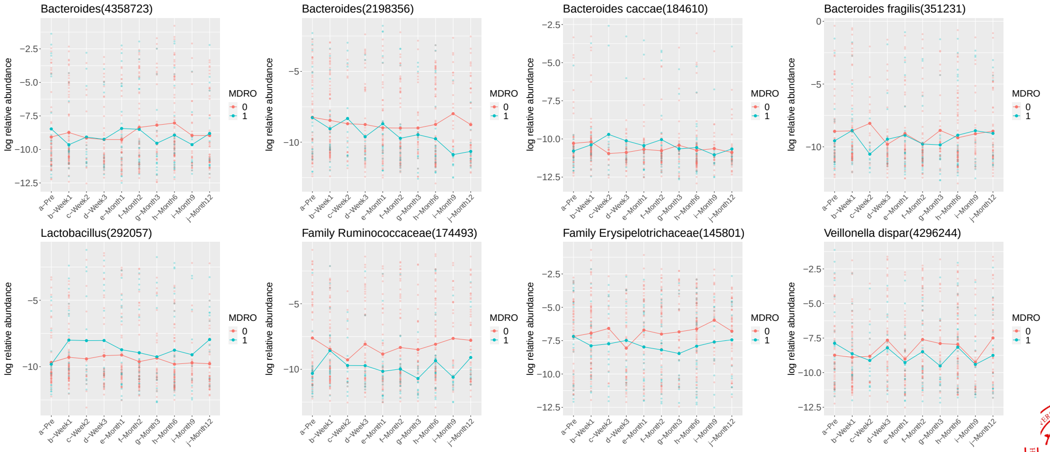
Outline

1. Introduction
2. Model setup
3. An iterative algorithm
4. Theoretical properties
5. Simulations
6. Real world application

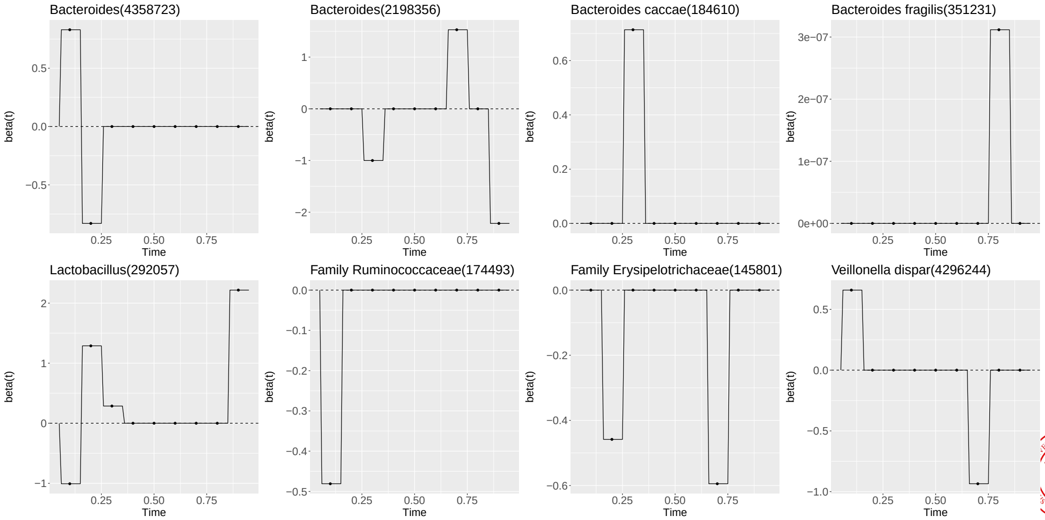
Colonizing MDRB and the intestinal microbiome

- Infections by Multidrug-resistant bacteria (MDRB) remain a leading cause of morbidity and mortality after liver transplantation(LT).
- Gut dysbiosis characteristic of end-stage liver disease may predispose patients to intestinal MDRB colonization and infectoin, in turn exacerbating dysbiosis.
- After quality control, data of 131 patients during one year after LT is collected. At Optional Taxonomic Units(OTU) level, 878 different taxons are identified.
- Colonizing status of MDRB for each patients is taken as response.

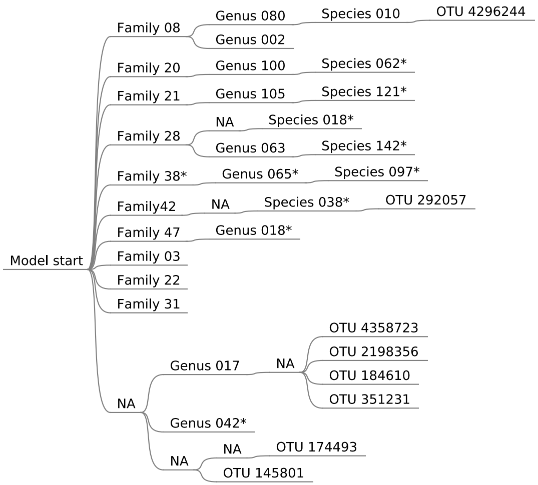
Variable selection



Estimation of functional parameters (after FLiRTI procedure)



Results at different taxonomic levels



Taxonomic levels: Family, Genus and Species.
NA means the corresponding taxon is not included in the model.

Asterisk means the current level is not officially recognized.

Future work

- Deal with confounders such as antibiotic treatment
- From observational study to causal inference
- Post-selection inference

Reference I

Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures [J/OL]. Biometrika, 1984, 71(2): 323-330. DOI: 10.1093/biomet/71.2.323.

Fan Y, James G M, Radchenko P. Functional additive regression [J/OL]. The Annals of Statistics, 2015, 43(5): 2296-2325. DOI: 10.1214/15-aos1346.

James G M, Wang J, Zhu J. Functional linear regression that's interpretable [J/OL]. The Annals of Statistics, 2009, 37(5A): 2083-2108. DOI: 10.1214/08-aos641.

Lin W, Shi P, Feng R, et al. Variable selection in regression with compositional covariates [J/OL]. Biometrika, 2014, 101(4): 785-797. DOI: 10.1093/biomet/asu031.

Ramsay J O, Silverman B W. Applied functional data analysis: Methods and case studies [M/OL]. Springer New York, 2002. DOI: 10.1007/b98886.

Redd A. A comment on the orthogonalization of b-spline basis functions and their derivatives [J/OL]. Statistics and Computing, 2011, 22(1): 251-257. DOI: 10.1007/s11222-010-9221-0.

Reference II

Sun Z, Xu W, Cong X, et al. Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome [J/OL]. The Annals of Applied Statistics, 2020, 14(3). DOI: 10.1214/20-aoas1357.

Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning problems [J/OL]. Statistics and Computing, 2014, 25(6): 1129-1141. DOI: 10.1007/s11222-014-9498-5.

Zhang X, Pei Y F, Zhang L, et al. Negative binomial mixed models for analyzing longitudinal microbiome data [J/OL]. Frontiers in Microbiology, 2018, 9. DOI: 10.3389/fmicb.2018.01683.