High Dimensional Count Data Analysis

Joshua D Habiger [1] [2]

Professor
Department of Statistics
Oklahoma State University

October 13, 2025

[1] Much of this based on Stat PhD dissertations by Emmanuel Asare, Huizi Wang. Many faculty and student collaborators in Plant Pathology and Entomology, Animal and Food Sciences at OSU

[2] Portion of Wang's dissertation supported by National Institute of Biosecurity and Microbial Forensics and Dept. Plant Pathology and Entomology Dept. at OSU.

Outline

1. **HD Count Data**
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - Assessment
3. HD Classifiction Analysis
   - Standard Methods
   - Proposed Methods
   - Assessment

Wheat Data: 16s rRNA technology

Sample	Biomass(g)	Bacteria 1	Bacteria 2	...	Bacteria 778
1	0.86	0	4	...	16
2	1.34	0	3	...	10
3	1.81	4	0	...	29
4	2.37	11	1	...	18
5	3.00	12	0	...	13

- Data[3]:
  - $N = 5$ wheat rhizosphere samples
  - $M = 778$ rhizobacteria per sample
- Research Question: Which Rhizobacteria are associated with biomass?
- **Basic Solution: Apply multiple hypothesis testing method**

[3] Anderson and Habiger (2012)

Meat Data: GC-MS technology

Sample	Color	Metabolite 1	Metabolite 2	...	Metabolite 171
1	ADC	92500	4055	...	7860
2	ADC	166598	4935	...	7323
⋮	⋮	⋮	⋮	⋱	⋮
11	RED	93647	5705	...	4340
12	RED	142929	4859	...	6680

- Data[4]:
  - N = 12 steaks (6 abnormally dark colored and 6 red)
  - M = 171 metabolites
- Research Question: Which metabolites are associated with meat color?
- **Basic Solution: Apply multiple hypothesis testing method**

[4] Kiyimba F, Cassens D, Hartson SD, Rogers J, Habiger J, Mafi GG, Ramanathan (2022)

Citrus Data: EDNA technology

Sample	Pathogen	Eprobe 1	Eprobe 2	...	Eprobe 8847
1	NO	0	0	...	0
2	NO	0	1	...	0
⋮	⋮	⋮	⋮	⋱	⋮
21	YES	1	1	...	0
22	YES	2	0	...	1

- Data[5]:
  - N = 22 Citrus samples (11 healthy and 11 pathogenic)
  - M = 8847 E-probes for Candidatus Liberibacter Asiaticus (CLAS) pathogen
- Research Question: Can we diagnose a new sample?
- **Basic Solution: Classification analysis**

[5] Dang T, Wang H, Espindola AS, Habiger J, Vidalakis G, Cardwell K (2023)

Other EDNA Data: Same question different pathogen

Pathogen Names	# E-probes (M)	# Samples (N)
Candidatus Liberibacter Asiaticus	8847	22
Citrus Tristeza Virus	823	15
Citrus Psorosis Virus	203	21
Citrus Virus A	151	21
Citrus Vein Enation Virus	90	19
Citrus Variegated Virus	86	22
Citrus Tatter Leaf Virus	51	20
Citrus Leaf Blotch Virus	27	17
Citrus Yellow Vein Associated Virus	24	22
Citrus Exocortis Viroid	6	18

Challenges and Solution Path

- In all settings of interest
  - High dimensional count data: $X_{N\times M}$ or $Y_{N\times M}$ ($M \gg N$)
  - Another variable (continuous or discrete) for testing or prediction
- Challenges we want to address **simultaneously**
  - Overdispersed
  - Zero-inflated
  - Heterogeneity (across samples and/or variables)
  - Overfitting
  - Computational complexity
  - Analytical results
- **Guiding principle**: Want solutions that will work well across a range of "omics" and beyond, i.e. FLEXIBLE

Outline

1. HD Count Data
2. **HD Multiple Testing**
   - Standard Methods
   - Proposed Methods
   - Assessment
3. HD Classification Analysis
   - Standard Methods
   - Proposed Methods
   - Assessment

Revisiting the Wheat Rhizosphere Data

- Background:
  - Rhizosphere: Area of the soil near roots
  - Rhizosphere microbiome: Microorganisms / bacteria in the rhizosphere
  - Basic wheat rhizosphere microbiome study: **Who's there / who's abundant**?
  - Idea: If we know who's there we can intervene
- Refined research question:
  - Are most abundant bacteria actually most relevant?
  - Is the abundance = association hypothesis true?
  - Compare micriobiome to **productivity associated** microbiome

Is Fred abundant? Is he associated with productivity?

"Fred's too lazy to fix things around the house. On the plus side, he's also too lazy to break things."

A Formal Problem Statement

x	Y_1	Y_2	...	Y_M
x_1	Y_{11}	Y_{12}	...	Y_{1N}
x_2	Y_{21}	Y_{22}	...	Y_{2N}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
x_N	Y_{N1}	Y_{N2}	...	Y_{NM}
Sufficient Stats:	(T_1, N_1)	(T_2, N_2)	...	(T_M, N_M)

- Data:
  - Y_{nm} abundance of mth bacteria in nth sample
  - x_n shoot biomass in nth sample (proxy for productivity)
- Model: $Y_{nm} \sim \text{Pois}(\mu_{nm})$ for $log(\mu_{nm}) = \alpha_m + \beta_m x_n$
- Hypotheses: $H_m : \beta_m = 0$
- Statistics: $(N_m, T_m) = (\sum_n Y_{nm}, \sum_n x_n Y_{nm})$ sufficient for (α_m, β_m)

Classical BH Procedure Approach

Step 1: For $m = 1, 2, .., M$ compute p-value P_m for each bacteria

- Compute Z-scores $Z_m = \frac{T_m - E_0[T_m|N_m=n_m]}{\sqrt{Var_0(T_m|N_m=n_m)}}$
- Get p-Values: $P_m = \Pr(|Z_m| \geq |z_m|)$ under $Y_m|N_m = n_m \sim Mult(n_m, \frac{1}{N}1)$

Step 2: Reject k nulls with smallest p values for $k = \max\{i : P_{(i)} \leq \alpha \frac{i}{M}\}$

Results: Discover 34 "productivity-associated" bacteria for $\alpha = 0.05$.

Empirical Bayes LFDR Approach (MODELING Z SCORES)

Step 1: Estimate local FDR statistics

- Mixture model: $Z_m \sim f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$
- Local FDR: $LFDR(z) = \frac{\pi_0 f_0(z)}{f(z)} = \Pr(H_m \text{ true } | Z_m = z)$
- Local FDR statistics: $LFDR_m = LFDR(Z_m)$
- Adaptive/Empirical Bayes: $\hat{\pi}_0, \hat{f}_1 \rightarrow \widehat{lFDR}_m$

Step 2: Reject k nulls with smallest $\widehat{LFDR}_m$ for $k = \max\left\{m : \sum_{i=1}^{m} \widehat{LFDR}_{(i)} \leq \alpha m\right\}$

Results: Discover 99 "productivity-associated" bacteria for $\alpha = 0.05$.

What is "significant"?

- Abundance: Bacteria m is found $Y_{1m} + ... + Y_{5m} = n_m$ times.
- Productivity-Associated Abundance: Large $|Y_{nm}/n_m - 0.2|$ or $|\hat{\beta}_m|$.
  - $E[Y_{nm}/n_m] = 0.2$ under $H_m : \beta_m = 0$

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{LFDR}(z_m)$	Discover
1	0.36	0.50	0.00	0.07	0.07	?	?	?	?
2	0.15	0.13	0.28	0.25	0.19	?	?	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	—	—

Which species is discovered?

What is "significant"?

- Abundance: Bacteria m is found $Y_{1m} + ... + Y_{5m} = n_m$ times.
- Productivity-Associated Abundance: Large $|Y_{nm}/n_m - 0.2|$ or $|\hat{\beta}_m|$.
  - $E[Y_{nm}/n_m] = 0.2$ under $H_m : \beta_m = 0$

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{LFDR}(z_m)$	Discover
1	0.36	0.50	0.00	0.07	0.07	**-1.09**	?	?	?
2	0.15	0.13	0.28	0.25	0.19	**0.19**	?	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	—	—

Which species is discovered?

What is "significant"?

- Abundance: Bacteria m is found $Y_{1m} + ... + Y_{5m} = n_m$ times.
- Productivity-Associated Abundance: Large $|Y_{nm}/n_m - 0.2|$ or $|\hat{\beta}_m|$.
  - $E[Y_{nm}/n_m] = 0.2$ under $H_m : \beta_m = 0$

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{LFDR}(z_m)$	Discover
1	0.36	0.50	0.00	0.07	0.07	**-1.09**	**11**	?	?
2	0.15	0.13	0.28	0.25	0.19	**0.19**	**911**	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	×

Which species is discovered?

What is "significant"

- Abundance: Bacteria m is found $Y_{1m} + ... + Y_{5m} = n_m$ times.
- Productivity-Associated Abundance: Large $|Y_{nm}/n_m - 0.2|$ or $|\hat{\beta}_m|$.
  - $E[Y_{nm}/n_m] = 0.2$ under $H_m : \beta_m = 0$

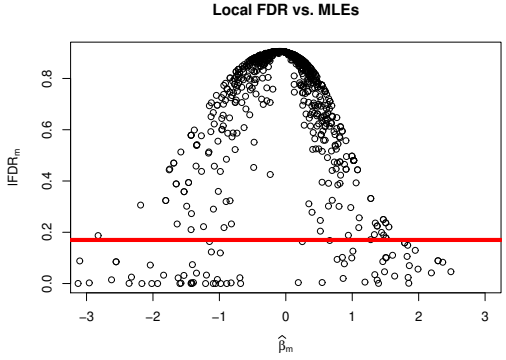
m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{LFDR}(z_m)$	Discover
1	0.36	0.50	0.00	0.07	0.07	**-1.09**	**11**	0.29	×
2	0.15	0.13	0.28	0.25	0.19	**0.19**	**911**	0.003	✓
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	×

The more abundant species is discovered and the most associated species is not!

Statistics Show Fred is Associated with Productivity ($P < .05$)!

"Fred's too lazy to fix things around the house. On the plus side, he's also too lazy to break things."

Its not a fluke



- Theorem 2 in Habiger, Watts Anderson (2017): $Lfdr(z_m), P(z_m) \to 0$ as $n_m \to \infty$ if $\beta_m \neq 0$.
- Remark: $7 \leq n_m \leq 911$

Outline

1. HD Count Data
2. HD Multiple Testing
   - Standard Methods
   - **Proposed Methods**
   - Assessment
3. HD Classification Analysis
   - Standard Methods
   - Proposed Methods
   - Assessment

Finite Mixture Model and Hypotheses

Main Idea: Model $\boldsymbol{Y}_m = (Y_{1m}, ..., Y_{Nm})$ (not Z_m or P_m!) and assume $\beta_m \in \{0, b_1, ..., b_J\}$

- Likelihood: $Y_{nm} \sim Poi(\mu_{nm})$ and $\log(\mu_{nm}) = \alpha_m + \beta_m x_n$
- Prior: $\Pr(\alpha_m = a_i, \beta_m = b_j) = \Pr(\alpha_m = a_i) \Pr(\beta_m = b_j) = \pi_{i.}\pi_{.j}$
  - $a_1, a_2, ..., a_I$ and $b_0, b_1, .., b_J$ unknown fixed except for known $b_0 = 0$
  - $\pi_{1.}, \pi_{2.}, ..., \pi_{I.}$ and $\pi_{.0}, \pi_{.1}, ..., \pi_{.J}$ unknown fixed
- Theoretical Null Hypothesis: $H_m : \beta_m = 0$
  - $\Pr(\beta_m = 0) = \pi_{.0}$
- (Towards an) "Empirical Null" Hypothesis: $H_m : \beta_m \in [-\epsilon, \epsilon]$
  - $\Pr(\beta_m \in [-\epsilon, \epsilon]) = \pi_{.0} + \sum_{j=1}^{J} \pi_{.j} I(b_j \in [-\epsilon, \epsilon])$

Proposed Local FDR Procedures

1. **Specify (possibly I), J, and ϵ in** $H_m : \beta_m \in [-\epsilon, \epsilon]$

2. Get MLE via:

$$l_Y(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\pi}) = \sum_m \log\left(\sum_{i,j} \pi_{i\cdot}\pi_{\cdot j} p(\boldsymbol{y}_m | a_i, b_j)\right) \text{ or } l_Y(\boldsymbol{b}, \boldsymbol{\pi}) = \sum_m \log\left(\sum_j \pi_{\cdot j} p(\boldsymbol{y}_m | b_j, n_m)\right)$$

3. Estimate the Oracle $LFDR_m$ or $CLFDR_m$: plug $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\pi}}$ in for

$$LFDR_m(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\pi}) = Pr(\beta_m \in [-\epsilon, \epsilon] | \boldsymbol{y}_m) = \frac{\sum_i \sum_{j:b_j \in [-\epsilon,\epsilon]} \pi_{i\cdot}\pi_{\cdot j} p(\boldsymbol{y}_m | a_i, b_j)}{\sum_i \sum_j \pi_{i\cdot}\pi_{\cdot j} p(\boldsymbol{y}_m | a_i, b_j)}$$

   or could use

$$CLFDR_m(\boldsymbol{b}, \boldsymbol{\pi}) = \Pr(\beta_m \in [-\epsilon, \epsilon] | \boldsymbol{y}_m, n_m) = \frac{\sum_{j:b_j \in [-\epsilon,\epsilon]} \pi_{\cdot j} p(\boldsymbol{y}_m | b_j, n_m)}{\sum_j \pi_{\cdot j} p(\boldsymbol{y}_m | b_j, n_m)}$$

4. Reject r null hypotheses with $r = \max\left\{j : \frac{1}{j}\sum_{i=1}^{j} \widehat{LFDR}_{(i)} \leq \alpha\right\}$ (or use $\widehat{CLFDR}_{(i)}$)

Outline

1. **HD Count Data**
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - **Assessment**
3. HD Classification Analysis
   - Standard Methods
   - Proposed Methods
   - Assessment

Now what is "significant"?

Question: Now which species is discovered?

Recall: LFDR procedure for $H_m : \beta_m = 0$

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{LFDR}_m$	Disc.
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.29	x
2	0.15	0.13	0.28	0.25	0.19	0.19	911	0.003	✓
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

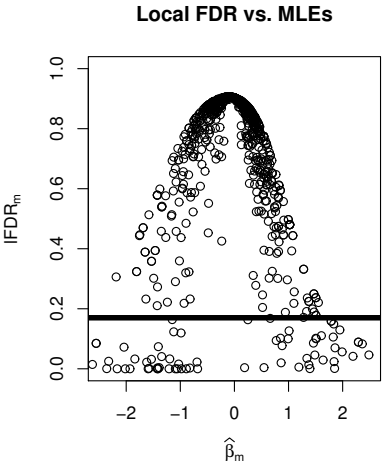
- Model: $Z_m \sim \pi_0 f_0(z) + (1-\pi_0) f_1(z)$

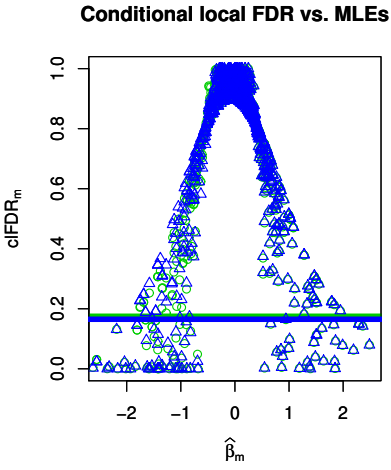
CLFDR Procedure for theoretical null $H_m : \beta_m = 0$

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{CLFDR}_m$	Disc.
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.10, 0.12	✓
2	0.15	0.13	0.28	0.25	0.19	0.19	911	1, 1	x

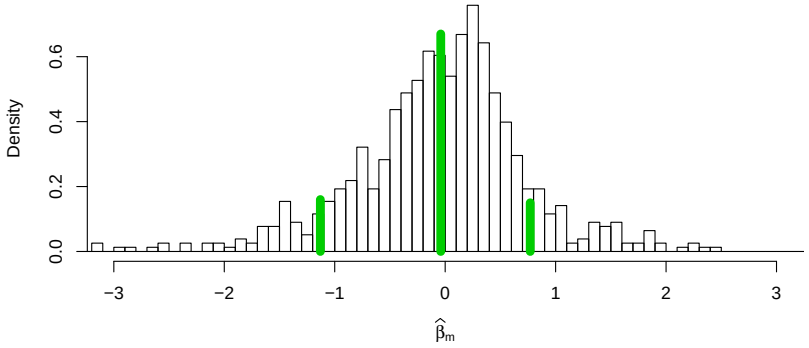
- Model: $\beta_m \in \{0, \hat{b}_1, \hat{b}_2\} = \{0, -1.1, 0.8\}$
- Model: $\beta_m \in \{0, \hat{b}_1, \hat{b}_2, \hat{b}_3\} = \{0, -2.7, -1.0, 0.8\}$

Is Fred Productive? Results inconclusive (P> .05)



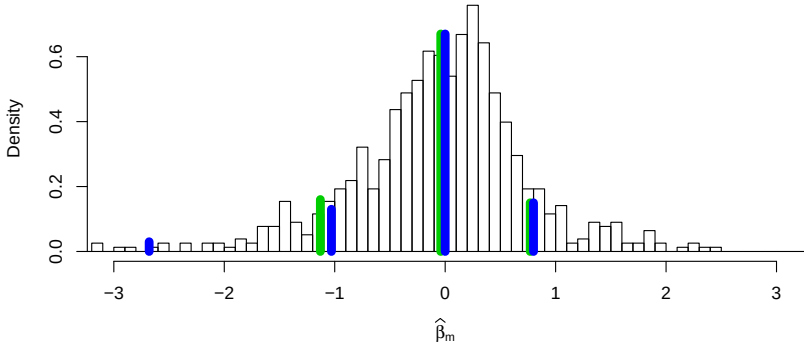


Empirical null safeguards against overfitting



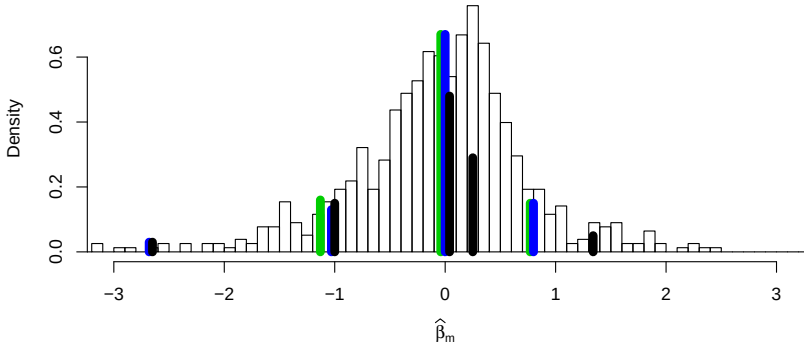
- $\Pr(\beta_m = 0) = \Pr(\beta_m \in [-\epsilon, \epsilon]) = 0.69$

Empirical null safeguards against overfitting



- $\Pr(\beta_m = 0) = \Pr(\beta_m \in [-\epsilon, \epsilon]) = 0.69$

Empirical null safeguards against overfitting



- $\Pr(\beta_m = 0) = 0.48$ but $\Pr(\beta_m \in [-\epsilon, \epsilon]) = 0.75$

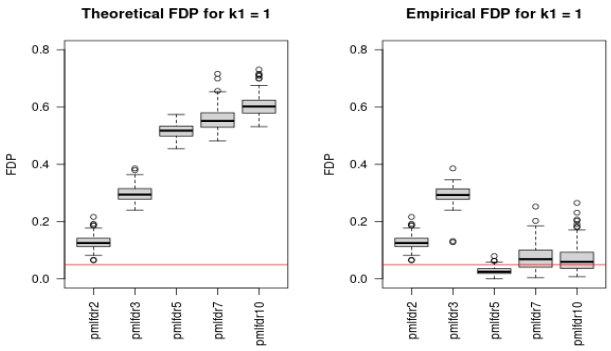
Use and Abuse of FMM

Practical setting simulation:

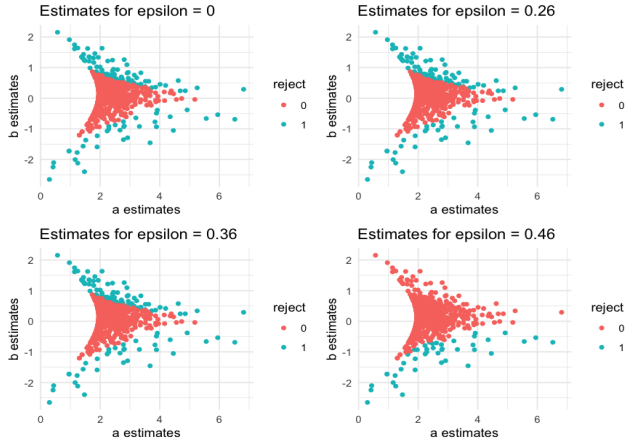
- $H_m : \beta_m \in [-.35, .35]$ since $log(2)/2 = 0.35$ (center and scale x!)
- $\beta_m = 0 + .35Un(-1, 1)$ wp $\pi_0 = 0.7$
- $\beta_m = 0.7 + .35Un(-1, 1)$ wp $1 - \pi_0 = 0.3$
- $J = 2, 3, 5, 7, 10$ alternative components

Do's and Don'ts

- **Do not** use many β_m components with theoretical null
- Do use many β_m components with thoughtful ϵ if using empirical null
- Do use few β_m components if using theoretical null

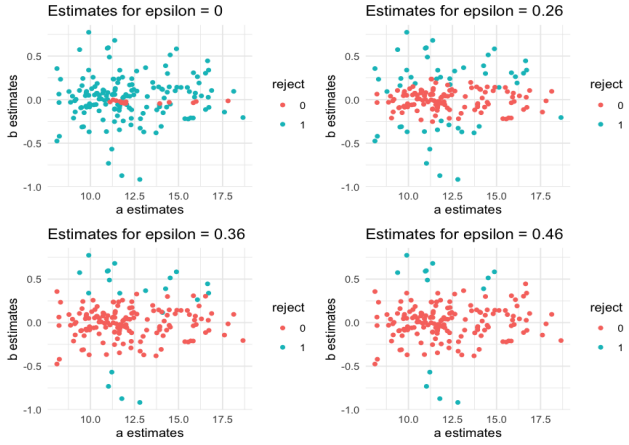


Wheat Data: Simple Model vs. ϵ



- Simple model (based on AIC, BIC): $\alpha_m \in \{a_1, ..., a_7\}$ and $\beta_m \in \{0, b_1, b_2\}$
- Choose ϵ sufficiently small!

Meat Data: Complex Model vs ϵ



- Complex model (based on AIC, BIC): $\alpha_m \in \{a_1, ..., a_{13}\}$ and $\beta_m \in \{0, b_1, ..., b_{10}\}$
- Choose ϵ sufficiently large!

Asymptotically Optimal FDR Control

Assuming model is correct with fixed ϵ, I, J and plugging in consistent estimators of $\boldsymbol{a}, \boldsymbol{b}, \pi$:

- $LFDR_m(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\pi}) \xrightarrow{p} LFDR_m(\boldsymbol{a}, \boldsymbol{b}, \pi)$ as $M \to \infty$
- Asymptotically valid (mFDR control) and asymptotic optimality (minimizes mFNR) procedure for FDR control as $M \to \infty$
- In general MLE invariance can give consistent estimators of other expressions, like **overdispersion** etc.

Overdispersion under the null

$$\frac{Var(Y_{nm}|\beta_m \in [-\epsilon, \epsilon])}{E[Y_{nm}|\beta_m \in [-\epsilon, \epsilon]]} = 1 + \phi(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\pi}, \epsilon, x_n) \geq 1$$

where $\phi(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\pi}, \epsilon, x_n)$

- $= 0$ only if $\epsilon = 0$ and $\boldsymbol{a} = a_1$ since $Y_{nm} \sim Pois(e^{a_1})$ under H_m
- > 0 (overdispersion) whenever
  - α_m is allowed to vary: ex. $\alpha_m \in \{a_1, a_2\}$
  - β_m is allowed to vary: ex. $H_m : \beta_m \in [-\epsilon, \epsilon]$

Concluding Remarks

Finite mixtures of log linear models for multiple testing does

- Address heterogeneity across variables via modeling $\beta_m \in \{0, b_1, ..., b_J\}$ (not Z_m or P_m)
- Allow for asymptotic analysis
- Is sufficiently flexible to
  - model overdispersion
  - model most data well
  - be misused too! (select ϵ or keep it simple)

Some open questions

- *CLFDR* mixtures of multinomials vs *LFDR* mixtures of Poissons?
- Model for $(\hat{\theta}_m, SE_m)$ vs $(\hat{\theta}_m/SE_m, W_m)$ vs (P_m, W_m) vs (P_m, E_m) vs and relevant optimality criteria
- Asymptotics for CLFDR
- Zero inflation (Pois(ϵ) vs δ_0)
- Across sample heterogeneity
  - Much work on normalizing samples.
  - Q: Include say $\log(S_n) = \log(NY_{n.}/Y_{..})$ (and vs. or) $\beta_m \in [-\epsilon, \epsilon]$?

More Open Questions

Sample	Type	Var	Dep	FGS	SW	GW	HW	Ker	B_1	B_2	...	B_{1000}
1	Still	Byrd	Bott	6	28.5	0.15	0.04	0	196	15	...	0
2	Still	Byrd	Mid	9	23	0.06	0.03	2	200	23	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
108	Still	Byrd	Top	9	33.75	0.18	0.07	2	0	0	...	0

- Research Question: Which bacterial counts (B_m) are associated with phenotype variables, in what way, and across what environments? [6]
- **Basic Solution:?**
  - Maybe $B_{mn} \sim Pois(\mu_{mn})$ with $\log(\mu_{mn}) = \alpha_m + \beta_m f(x_{1n}, .., x_{qn})$
  - What is f? Inference for f?
  - Replication across environments?

[6] Data from D. Ramos STAT 5063 Statistical Machine Learning project and Plant Pathology MS Thesis. PI A. Espindola

Outline

1. HD Count Data
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - Assessment
3. **HD Classification Analysis**
   - Standard Methods
   - Proposed Methods
   - Assessment

Revisiting EDNA Data

- Data Generation:
  1. Create "E-probes" for a pathogen of interest (20 - 40 base pairs each)
  2. Measure number of hits per E-probe in about 10 control (healthy) samples and 10 treatment (infected) samples
- Objective: Train a classifier that can input new biological sample and determine if sample infected
- **Refined Objective**: Want classifier that can be hard-coded in Microbe Finder™ (MiFi) software:
  - Want closed form stable solution!
  - Eliminate redundant / unnecessary Eprobes (computational and human hour costs)
  - Handle over-dispersion and zero inflation
  - High classification power

Outline

1. HD Count Data
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - Assessment
3. HD Classification Analysis
   - **Standard Methods**
   - Proposed Methods
   - Assessment

Modeling Approaches

Sample	y	x_1	x_2	...	x_p
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	...	x_{np}
New	?	X_1	X_2	...	X_p

- Data: $X = (X_1, ..., X_p)$ are counts and $Y \in \{0, 1\}$
- Multivariate Model Classifier: $\hat{Y} = 1$ if $\frac{p(x_1,...,x_p|y=1)}{p(x_1,...,x_p|y=0)} > K$.
  - Assume $X_j|Y = y \sim ZINB\left(\pi_j(y), \mu_j(y), \phi_j(y)\right)$ or many others
- Logistic Reg Classifier: $\hat{Y} = 1$ if $\text{logit}(\Pr(Y = 1|x_1, ..., x_p)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p > K$
- Get penalized MLE's (not closed form) and plug into classifier.

Outline

1. HD Count Data
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - Assessment
3. HD Classifiction Analysis
   - Standard Methods
   - **Proposed Methods**
   - Assessment

Main Idea Motivation

- We want 1. closed form, 2.penalized parameter estimation that 3. models overdispersion and zero-inflation and 4. selects variables

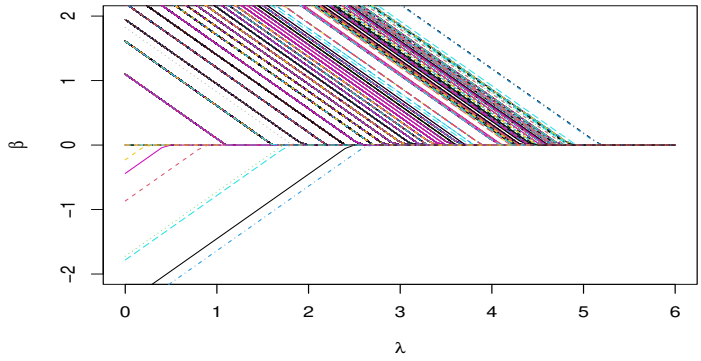
- Towards main idea: Observe logistic reg and Poisson LDA classifiers are the same:

$$C(x) = \log\left(\frac{\prod_j p(x_j|Y=1)}{\prod_j p(x_j|Y=0)}\right) \propto [\log(\mu_1(1)/\mu_1(0))]x_1 + ... + [\log(\mu_p(1)/\mu_p(0))]x_p$$

$$C(x) = \text{logit}(\Pr(Y=1|x_1,...,x_p)) \propto \beta_1 x_1 + ... + \beta_p x_p$$

- Penalized method of moment estimators
  1. $\hat{\beta}_j = \log(\bar{x}_j(1)/\bar{x}_j(0))$
  2. $\hat{\beta}_j^{\lambda} \leftarrow S_\lambda(\hat{\beta}_j) = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$

Soft Thresholding Illustration on CLAS data



- $\hat{\beta}_j^{\lambda}$ vs λ: Observe λ shrinks MOM estimates towards 0 + selects variables
- About 3 lines of R code!

Zero inflation and over dispersion

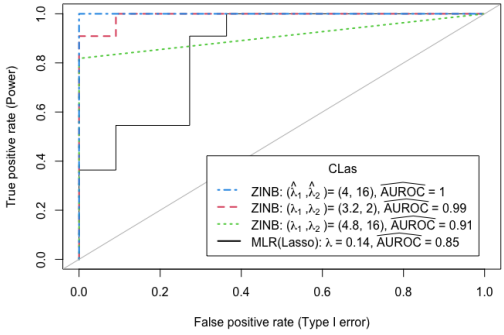
$$C(x) = \beta_1 x_1 + ... + \beta_p x_p + \alpha_1 \delta_0(x_1) + ... + \alpha_p \delta_0(x_p)$$

1. Use MOM estimates of $\pi_j(y), \mu_j(y), \phi_j$ to get MOM estimates of β_j, α_j
   - Can use any model from Pois to ZINB for initial estimates
   - Must include $\delta_0(x_j)$ if want zero inflation
2. Get $\hat{\beta}_j^{\lambda_1} = S_{\lambda_1}(\hat{\beta}_j)$ and $\hat{\alpha}_j^{\lambda_2} = S_{\lambda_2}(\hat{\alpha}_j)$
3. Plug $\hat{\beta}_j^{\lambda_1}$ and $\hat{\alpha}_j^{\lambda_2}$ into $C(x)$ to get

Outline

1. HD Count Data
2. HD Multiple Testing
   - Standard Methods
   - Proposed Methods
   - Assessment
3. HD Classification Analysis
   - Standard Methods
   - Proposed Methods
   - **Assessment**

Zero Inflated Negative Binomial vs Logistic Regression



- Many choices for λ_1, λ_2 (and models!) yield perfect group separation
- Logistic regression model can't separate groups
- Q: For $\hat{\lambda}_1$ and $\hat{\lambda}_2$ estimated via CV, what is the test AUROC?

Concluding Remarks

Penalized MOM estimation for HD classification

- Is sufficiently flexible to
  - model overdispersion
  - model zero inflation
  - to allow for high classification power
  - to be misused (selectively choose λs)
- Is sufficiently simple (closed form)
  - to be easily automated
  - to be misused
- Performs variable selection (E probe selection)
- Does not select *the* correct model

Some open questions

- other penalties and tuning parameter selection
- other moments
- asymptotics

Some more open questions

Pathogen Names	# E-probes (M)	# Samples (N)
Candidatus Liberibacter Asiaticus	8847	22
Citrus Tristeza Virus	823	15
Citrus Psorosis Virus	203	21
Citrus Virus A	151	21
Citrus Vein Enation Virus	90	19
Citrus Variegated Virus	86	22
Citrus Tatter Leaf Virus	51	20
Citrus Leaf Blotch Virus	27	17
Citrus Yellow Vein Associated Virus	24	22
Citrus Exocortis Viroid	6	18

- Assessment across all data sets?
- Pooling information across data?
- More efficient computation still?

Thanks to Collaborators

HD classification analysis methods and data

- Wang, H. (2025). Efficient Classification Methods for Sparse High-Dimensional Count Data with Model Selection as Motivated by Microbe Finder. Doctoral dissertation research, Department of Statistics, OSU.
- Dang, T., Wang, H., Espindola, A., **Habiger, J.**, Vidalakis, G., Cardwell, K. Cardwel (2023). Development and statistical validation of e-probe diagnostic nucleic acid analysis (EDNA) detection assays for the detection of citrus pathogens from raw high throughput sequencing data. *PhytoFrontiers* 3(1): 113-123.

Multiple testing methods and data

- Asare, E. (2025). Multiple Testing Procedures for Count Data Based on the Local False Discovery Rate. Doctoral dissertation, Department of Statistics, OSU.
- Frank, K., Ramanathan, R. Hartson, S., Rogers, J. Habiger, J, Mafi G (2023). Integrative proteomics and metabolomics profiling to understand the biochemical basis of beef muscle darkening at a slightly elevated pH. *Journal of Animal Science* (101).
- Fornah, A., Anderson, M. and Habiger, J. (2021). The Relationship Between Rhizobacteria Auxin Production and Wheat Biomass is Negative. *Communications in Soil Science and Plant Analysis*, 52(9), 985-997.
- Habiger, J. (2019). Discussion of "Covariate-Assisted Ranking and Screening for Large-scale Two-sample Inference by Cai, Sun and Wang." *Journal of the Royal Statistical Society Series B,* 81(2), 226-227.
- Habiger, J., Watts, D. and Anderson, M. (2017). Multiple Testing with Heterogeneous Multinomial Distributions. *Biometrics*, 73(2), 562-570.
- Anderson M. and Habiger J. (2012). Characterization and Identification of Productivity Associated Rhizobacteria in Wheat. *Applied and Environmental Microbiology*, 78(12), 4434-4446.

Thanks to Emmanuel Asare, Huizi Wang, Pratyaydipta Rudra, Ye Liang, Lan Zhu, Ranjith Ramanthan, Kitty Cardwell, Michael Anderson, Andres Espindola et al.