A Two-Stage Kernel Machine Regression Model for Integrative Analysis of Alpha Diversity

Runzhe Li

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

March 23, 2020

Content

1. Introducution
   - Human microbiota and microbiome
   - Integrative analysis
   - Alpha diversity
2. Methods
   - Kernel regression models
   - Association testing
3. Results
   - Simulation studies
   - Simulation results
4. Summary

Content

1. Introducution
    - Human microbiota and microbiome
    - Integrative analysis
    - Alpha diversity
2. Methods
    - Kernel regression models
    - Association testing
3. Results
    - Simulation studies
    - Simulation results
4. Summary

Introduction: Human microbiota and microbiome

- Human microbiota: ecological communities of microorganisms that reside in and on human body.
- Human microbiome: the collective genomes of resident microorganisms.

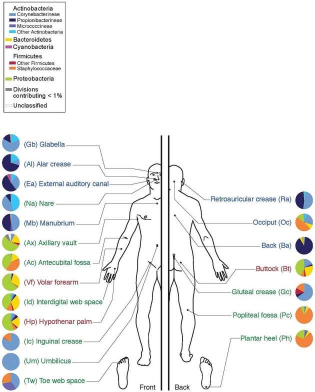
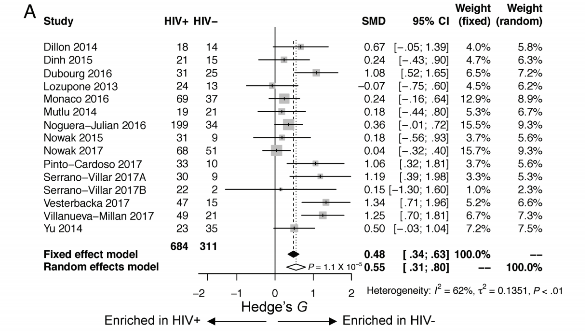


Figure: human skin microbiota [1]

[1] https://en.wikipedia.org/wiki/Microbiota

Introduction: Integrative Analysis

- The study to examine α-diversity measures between HIV^{-} and HIV^{+} individuals. 22 studies were identified with 17 datasets available for analysis, yielding 1032 samples. [2]



Study	HIV+	HIV−	SMD	95% CI	Weight (fixed)	Weight (random)
Dillon 2014	18	14	0.67	[−.05; 1.39]	4.0%	5.8%
Dinh 2015	21	15	0.24	[−.43; .90]	4.7%	6.3%
Dubourg 2016	31	25	1.08	[.52; 1.65]	6.5%	7.2%
Lozupone 2013	24	13	−0.07	[−.75; .60]	4.5%	6.2%
Monaco 2016	69	37	0.24	[−.16; .64]	12.9%	8.9%
Mutlu 2014	19	21	0.18	[−.44; .80]	5.3%	6.7%
Noguera−Julian 2016	199	34	0.36	[−.01; .72]	15.5%	9.3%
Nowak 2015	31	9	0.18	[−.56; .93]	3.7%	5.6%
Nowak 2017	68	51	0.04	[−.32; .40]	15.7%	9.3%
Pinto−Cardoso 2017	33	10	1.06	[.32; 1.81]	3.7%	5.6%
Serrano−Villar 2017A	30	9	1.19	[.39; 1.98]	3.3%	5.3%
Serrano−Villar 2017B	22	2	0.15	[−1.30; 1.60]	1.0%	2.3%
Vesterbacka 2017	47	15	1.34	[.71; 1.96]	5.2%	6.6%
Villanueva−Millan 2017	49	21	1.25	[.70; 1.81]	6.7%	7.3%
Yu 2014	23	35	0.50	[−.03; 1.04]	7.2%	7.5%
	684	311				
Fixed effect model			**0.48**	**[.34; .63]**	**100.0%**	**--**
Random effects model			**0.55**	**[.31; .80]**	**--**	**100.0%**

Heterogeneity: $I^2 = 62\%$, $\tau^2 = 0.1351$, $P < .01$

Figure: Figures from a study to investigate associations between gut microbial α-diversity and HIV status

[2] Tuddenham, S. A., Koay, W. L. A., Zhao, N., White, J. R., Ghanem, K. G., and Sears, C. L. (2020). The Impact of Human Immunodeficiency Virus Infection on Gut Microbiota α-Diversity: An Individual-level Meta-analysis. Clinical Infectious

Introduction: Integrative Analysis

- Analyzing data from **individual** study
  - → small sample size
  - → does not account for study heterogeneity
  - → inconsistent results due to the technical variability.
- Developing Integrative analysis from **multiple** studies
  - → address the potential biases
  - → boost statistical power and recover signals

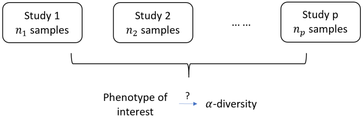


Figure: Integrative analysis leveraging information from multiple studies

Introduction: Alpha diversity

- Alpha diversity: summarizes the diversity **within** an ecological community
  - Species Richness: *how many?*
  - Species Evenness: *how different?*

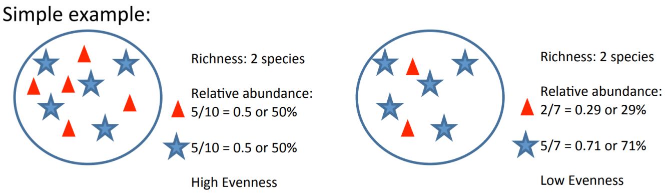


Figure: A toy example of species richness and evenness [3]

[3] http://users.unimi.it/dmora/materiali/ANU2012/BioInfoTools.pdf

Introduction: Questions of interest

- Our goal: identify the association between phenotype of interest and alpha diversity via integrative analysis of multiple studies.
- different hypothesis testings:
  - common effect test
  - heterogeneity test
  - joint test

Content

1. Introducution
   - Human microbiota and microbiome
   - Integrative analysis
   - Alpha diversity

2. Methods
   - Kernel regression models
   - Association testing

3. Results
   - Simulation studies
   - Simulation results

4. Summary

Methods: Two-Stage Kernel Machine Regression Model

Stage One

The association between alpha-diversity and HIV status via a linear mixed model

$$
\begin{aligned}
y_{ij} &= x_{ij}\beta_i + z_{ij}^\top \gamma + h_i + \epsilon_{ij} \\
Y_{N\times 1} &= X_{N\times p}\beta_{p\times 1} + Z_{N\times q}\gamma_{q\times 1} + h_{N\times 1} + \epsilon_{N\times 1}
\end{aligned}
$$

where $\beta(p \times 1), \gamma(q \times 1)$ are regression coefficients for fixed effects. $h_i \sim N(0, \sigma_h^2)$ are study-specific random effect which captures the difference between studies, $\epsilon_{ij} \sim N(0, \sigma_e^2)$ are error terms.

Methods: Two-Stage Kernel Machine Regression Model

Stage Two

We allow β_i to vary according to the study-specific characteristics

$$\beta_i = \beta_0 + f(G_i)$$

where $f(\cdot)$ is a function in reproducing kernel Hilbert space generated by a positive semidefinite kernel function $K(\cdot,\cdot)$. For example, a linear kernel $K(G_i, G_{i'}) = \sum_{j=1}^{r} G_{ij} G_{i'j}$ could measure the similarity between study i and study i'.

Methods: Association testing

Consider a special case where the second stage is:

$$\beta = \beta_0 \cdot \mathbf{1_p} + G\alpha, \alpha \sim N(0, \tau^2 I_{r\times r})$$

where $\beta \sim N(\beta_0 \cdot \mathbf{1_p}, \tau^2 K)$ and $K = GG'$.

Methods: Association Testing

Plug in and we can derive:

$$\begin{aligned} &Y = (X1_p\beta_0 + Z\gamma) + (XG\alpha + h) + \epsilon \\ &\alpha \sim N(0, \tau^2 I) \end{aligned}$$

We are interested in the following hypothesis testings.

1. (test for common effect) test $\beta_0 = 0$ under $\tau^2 = 0$
2. (test for heterogeneity) test $\tau^2 = 0$ under $\beta_0 = 0$
3. (test for heterogeneity) test $\tau^2 = 0$ without constraint on β_0
4. (joint test) combine test 1 and 3
5. (test for common effect) test $\beta_0 = 0$ without constraint on τ

Methods: Association Testing

1. (Burden test) test $\beta_0 = 0$ under $\tau^2 = 0$

$$Q_{\beta_0} = (Y - Z\hat{\gamma})^\top \hat{\Sigma_0}^{-1} X 1_p 1_p^\top X^\top \hat{\Sigma_0}^{-1} (Y - Z\hat{\gamma})$$

where the null MLE: $\hat{\gamma}, \hat{\sigma_h^2}, \hat{\sigma_e^2}$, and estimated null covariance matrix: $\hat{\Sigma_0} = \hat{\sigma_h^2} H + \hat{\sigma_e^2} I_N$ could be derived under the null model $Y = Z\gamma + h + \epsilon$.

We can further show that Q_{β_0} follows a scaled chi-square distribution with freedom 1.

Methods: Association Testing

2. (SKAT test) test $\tau^2 = 0$ under $\beta_0 = 0$

$$Q_{\tau_0^2} = (y - Z\hat{\gamma})^\top \hat{\Sigma_0}^{-1} XKX^\top \hat{\Sigma_0}^{-1} (Y - Z\hat{\gamma})$$

where $\hat{\gamma}$ and $\hat{\Sigma_0}$ are still derived under the null model $Y = Z\gamma + h + \epsilon$.

According to Davies method [4], Q_{τ^2} follows a mixture of chi-square distribution.

[4] Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(3), 323-333.

Methods: Association Testing

3. (Unconstrained SKAT test) test $\tau^2 = 0$ without constraint on β_0

$$Q_{\tau^2} = (y - \hat{\mu})^\top \hat{\Sigma}^{-1} X K X^\top \hat{\Sigma}^{-1} (y - \hat{\mu})$$

Let $\hat{\mu}$ denote the fitted value, $\hat{\Sigma}$ be the estimated covariance matrix under the model $Y = X 1_p \beta_0 + Z\gamma + \epsilon$.

Similarly, we could also show Q_{τ^2} follows a mixture of chi-square distribution.

Methods: Association Testing

4. (Optimal Joint Test) test $\beta_0 = 0$ and $\tau^2 = 0$

Let $Q_\rho = \rho Q_{\beta_0} + (1-\rho) Q_{\tau^2}$, where Q_{β_0} is derived from burden test, and Q_{τ^2} is derived from unconstrained SKAT test. Let p_ρ denote the p-value of Q_ρ. Then the test statistics is

$$T = \min_{0 \le \rho \le 1} p_\rho$$

Methods: Association Testing

5. (Unconstrained burden test) test $\beta_0 = 0$ without constraint on τ

$$Q_\beta = (Y - Z\hat{\gamma})^\top \hat{\Sigma}_\tau^{-1} X 1_p 1_p^\top X^\top \hat{\Sigma}_\tau^{-1} (Y - Z\hat{\gamma})$$

where $\hat{\Sigma}_\tau = \hat{\sigma_h^2} H + \hat{\sigma_e^2} I_N + \hat{\tau}^2 X K X^\top$ could be derived from the model $Y = Z\gamma + XG\alpha + h + \epsilon$.

Q_β follows a scaled chi-square distribution with freedom 1.

Content

1. Introducution
   - Human microbiota and microbiome
   - Integrative analysis
   - Alpha diversity
2. Methods
   - Kernel regression models
   - Association testing
3. Results
   - Simulation studies
   - Simulation results
4. Summary

Results: Simulation Studies

$$y_{ij} = x_{ij}\beta_i + z_{ij}^\top\gamma + h_i + \epsilon_{ij}$$
$$\beta_i = \beta_0 + G_i^\top\alpha, \alpha \sim N(0, \tau^2 I)$$

- Design matrix:
  - X_{ij}: HIV status $\{0, 1\}$
  - Z_{ij}: intercept, MSM $\{0, 1\}$, gender $\{0, 1\}$
  - $G_{p\times r}$: categorical $\{0, 1, 2, 3\}$: primer,sequence, DNA extraction,batch effect
- Parameters
  - $h \sim N(0, \sigma_h^2)$, $\epsilon \sim N(0, \sigma_e^2)$
  - change ratio of h and ϵ
- Sample size and dimensions:
  - # of study: p
  - # of microbiome characteristic: r

Results: Simulation Studies

$$y_{ij} = x_{ij}\beta_i + z_{ij}^\top \gamma + h_i + \epsilon_{ij}$$
$$\beta_i = \beta_0 + G_i^\top \alpha, \alpha \sim N(0, \tau^2 I)$$

- Type I error and power analysis
    - Type I error: $\beta_0 = 0$ and $\tau = 0$
    - Power case I: $\beta_0 \neq 0$ and $\tau = 0$
    - Power case II: $\beta_0 = 0$ and $\tau \neq 0$
    - Power case III: $\beta_0 \neq 0$ and $\tau \neq 0$

Results: Simulation Results

Type I error

	Empirical Type I Error	
level	0.05	0.01
LMM [†]	0.048	0.011
Burden	0.048	0.011
SKAT	0.054	0.006
Unconstrained SKAT	0.045	0.009
Optimal joint test	0.050	0.0095
Unconstrained burden	0.013	0.001

Table: Empirical type I error under 2000 simulations: $r = 3$, $p = 50$, $\sigma_h = 1$, $\sigma_e = 0.5$

[†] the default p-value of linear mixed model only testing for common effect in R.

Results: Simulation Results

Power

Case I : $\beta_0 = 0.1$ and $\tau = 0$

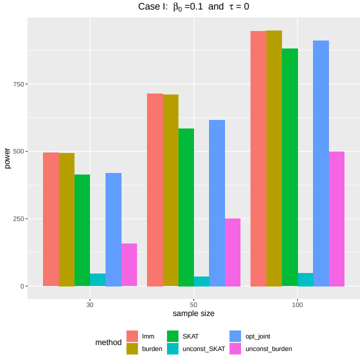


Figure: Power under simulation setting 1

Findings

- sample size increases → power increases
- overall performance: LMM, burden $>$ optimal joint test $>$ SKAT
- unconstrained SKAT test has power $\approx$ type I error

Results: Simulation Results

Power

Case II : $\beta_0 = 0$ and $\tau = 0.1$

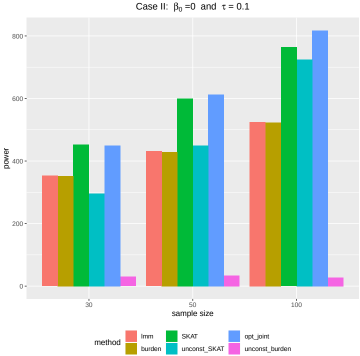


Figure: Power under simulation setting 2

Findings

- overall performance: optimal joint test,SKAT $>$ unconstrained SKAT $>$ LMM, burden
- unconstrained burden test has power $\approx$ type I error

Results: Simulation Results

Power

Case III : $\beta_0 = 0.1$ and $\tau = 0.1$

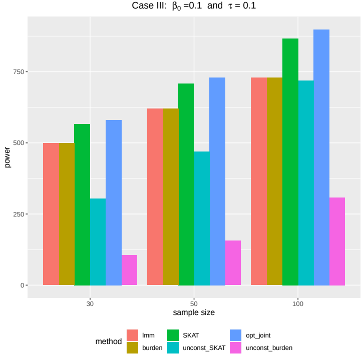


Figure: Power under simulation setting 3

Findings

- overall performance: optimal joint test is the most powerful test compared to other burden and variance component test.

Content

1. Introducution
   - Human microbiota and microbiome
   - Integrative analysis
   - Alpha diversity
2. Methods
   - Kernel regression models
   - Association testing
3. Results
   - Simulation studies
   - Simulation results
4. Summary

Summary

Conclusions

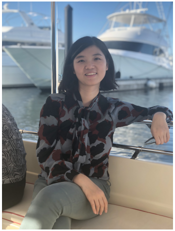
- propose a two-stage kernel machine regression model to associate alpha diversity with the phenotype of interests.
  - Stage one: model the relationship between the alpha diversity and the phenotype via a linear mixed model.
  - Stage two: incorporate the study-specific characteristics through a nonparametric function to allow for the between-study heterogeneity.
- construct several association testing problems.
- design the simulation studies.

Future work

- permutation tests for small sample size adjustment.
- application on real HIV studies.

Acknowledgement

This is joint work with Dr.Ni Zhao.



Ni Zhao

Reference I

Susan A Tuddenham, Wei Li A Koay, Ni Zhao, James R White, Khalil G Ghanem, and Cynthia L Sears.
The impact of human immunodeficiency virus infection on gut microbiota α-diversity: An individual-level meta-analysis.
Clinical Infectious Diseases, 70(4):615–627, 2020.

Amy D Willis.
Rarefaction, alpha diversity, and statistics.
Frontiers in microbiology, 10:2407, 2019.

Hyunwook Koh.
An adaptive microbiome α-diversity-based association analysis method.
Scientific reports, 8(1):1–12, 2018.

Han Chen, Jennifer E Huffman, Jennifer A Brody, Chaolong Wang, Seunggeun Lee, Zilin Li, Stephanie M Gogarten, Tamar Sofer, Lawrence F Bielak, Joshua C Bis, et al.
Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies.
The American Journal of Human Genetics, 104(2):260–274, 2019.

Zhong Wang, Ke Xu, Xinyu Zhang, Xiaowei Wu, and Zuoheng Wang.
Longitudinal snp-set association analysis of quantitative phenotypes.
Genetic epidemiology, 41(1):81–93, 2017.

Reference II

Yu-Ru Su, Chongzhi Di, Stephanie Bien, Licai Huang, Xinyuan Dong, Goncalo Abecasis, Sonja Berndt, Stephane Bezieau, Hermann Brenner, Bette Caan, et al.
A mixed-effects model for powerful association tests in integrative functional genomics.
The American Journal of Human Genetics, 102(5):904–919, 2018.

Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu.
Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test.
The American Journal of Human Genetics, 96(5):797–807, 2015.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin.
Rare-variant association testing for sequencing data with the sequence kernel association test.
The American Journal of Human Genetics, 89(1):82–93, 2011.

Seunggeun Lee, Michael C Wu, and Xihong Lin.
Optimal tests for rare variant effects in sequencing association studies.
Biostatistics, 13(4):762–775, 2012.

A Two-Stage Kernel Machine Regression Model for Integrative Analysis of Alpha Diversity

Runzhe Li

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

March 23, 2020