The Pennsylvania State University
The Graduate School

ERROR IN SCALE RELIANT COVARIATION ANALYSIS

A Thesis in
Informatics
by
Zhao Ma

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2023

The thesis of Zhao Ma was reviewed and approved* by the following:

Justin Silverman
Assistant Professor of Information Sciences and Technology
Thesis Advisor

Sharon Huang
Professor of Information Sciences and Technology

Jeffrey Bardzell
Associate Dean of Undergraduate and Graduate Studies
Professor of Information Sciences and Technology

*Signatures are on file in the Graduate School.

Abstract

People may be curious about how the false assumption of scale will affect the correlation estimates. A good way to explore this impact is to separate the covariation equation into compositional and scale parts, and see how the error in scale parts caused by erroneous assumption propagates to the final estimand. This thesis explores the impact of erroneous modeling assumptions regarding scale on the fidelity of covariation analysis estimates using high-throughput genomic data. We develop a mathematical representation of covariation analysis in terms of composition and scale and then use this foundation to study how errors in unmeasured scale parts propagate into correlation estimates. This study allows us to create new tools which allow applied researchers to characterize the sensitivity of correlation estimates to model misspecification and thereby identify when putative associations are likely false positives secondary to erroneous modeling assumptions.

Table of Contents

List of Symbols

$Cov(X, Y)$	The covariance of X and Y.
$X_{d\cdot}$	The d-th row of a matrix X.
$\rho(X, Y)$	The correlation of X and Y.
σ_X	The standard deviation of X.
$\mathbf{1}$	A vector whose elements are all 1.
$\mathbf{1}_{D\times D}$	A $D \times D$ matrix whose elements are all 1.
X^T	The transpose of amatrix X.
$\circ$	Hadamard Product.

Acknowledgments

I would like to thank my advisor Justin D Silverman who has provided me a lot of help through my master study. I am especially appreciate for your patience and outstanding guidance. Thanks to you, I not only learned knowledge but also feel the happiness of doing research. It is really my fortune to have you as my advisor.

Finally, I would like to thank my family and friends. Their constant support helps me overcome various difficulty throughout my master study.

Chapter 1
Introduction

High throughput genomic surveys, such as 16S rRNA gene profiling, play an important role in modern biomedical research [8]. Data from genomic surveys are often used to estimate correlations and covariation between taxa in microbial communities. For example, letting W_{dn} denote the abundance of microbial taxon $d \in \{1, \ldots, D\}$ in community $n \in \{1, \ldots, N\}$, researchers are often interested in estimating quantities of the form:

$$\Sigma_{d_1 d_2} = \mathrm{Cov}(\log W_{d_1 \cdot}, \log W_{d_2 \cdot}). \tag{1.1}$$

Such covariance estimates can answer fundamental and applied biomedical questions. For example, an outstanding problem in the study of microbiota is to infer interaction networks between bacteria in the hopes of then modulating microbial community structure for therapeutic benefit [1] [3] [6]. While there are a wide variety of methods that have been proposed to estimate pairwise correlations between microbial taxa [2] [10], recent research has shown that existing methods may be more limited than originally thought due to fundamental limitations of the measurement process used to obtain genomic survey data.

Above we have expressed covariation analysis in terms of the true abundances W, yet in practice researchers only have access to measurements of W which we can denote Y. Like W, these measurements can be expressed as tabular data with elements Y_{dn} denoting the number of DNA molecules observed (sequenced) that map to taxon d in the sample from community n. While there are a number of ways in which Y can differ from W including systematic bias from PCR Amplification [13] and measurement noise from the survey process, there are also fundamental limitations on the information content of these data. In particular, it is increasingly recognized that the scale of these data (defined as $Y_n^{\perp} = \sum_d Y_{dn}$) is typically driven by technical factors unrelated to the actual

scale (i.e., biological load) in the communities being studied ($W_n^{\perp} = \sum_d W_{dn}$). This lack of scale information has led a number of authors to call these data compositional, emphasizing that the only information contained in these data pertains to the relative (or proportional) abundance of the various taxa ($W_{dn}^{\parallel} = W_{dn}/W_n^{\perp}$).

Without scale information, estimation of many biomedically relevant quantities is ill-posed: knowledge of the system composition alone ($W^{\parallel}$) is insufficient to uniquely determine correlation or covariation between taxa [5]. There have been numerous tools proposed recently that attempt to address this problem with covariation analysis using various types of penalized estimation or data normalization [4] [5] [7] [11] [12]. Yet Nixon et AL [9] recently showed that such approaches carry fundamental limitations. They proved that any point estimator of inter-taxa correlations (e.g., estimator of $\Sigma_{d_1 d_2}$) must make strong and potentially implicit assumptions about the unmeasured system scale $W^{\perp}$, assumptions that could not be validated in practice but can dominate estimates [9]. For context, consider that penalized estimation is often studied in the setting of $p >> n$ problems, where there exist a finite number of model parameters that exceed the number of data points. In these situations penalization acts as prior knowledge regularizing estimates in finite data regimes. Much of the theory of penalized estimation relies on the fact that asymptotically, as $n \to \infty$, the estimator is consistent and any bias introduced by the penalization is dwarfed by the information content of the observed data. Yet with scale limitations such asymptotic behavior is not present. Even asymptotically these data fundamentally lack scale information and estimation problem remains ill-posed. As a result any bias introduced through the penalization remains even in the asymptotic limit [9].

People may be curious about how the false assumption of scale will affect the correlation estimates. A good way to explore the impact is to separate the equation 1.1 into compositional and scale parts, and then see how the error in scale parts caused by erroneous assumption propagates to the final estimand. Therefore, this thesis explores the impact of erroneous modeling assumptions regarding scale on the fidelity of covariation analysis estimates using high-throughput genomic data. Based upon the work of Nixon et al [9], we create a mathematical foundation for studying modeling assumptions in correlation analysis and then use this foundation to study how errors in modeling assumptions propagate into covariation estimates. This study allows us to create new tools which allow applied researchers to characterize the sensitivity of correlation estimates to model misspecification and thereby identify when putative associations are likely false positives secondary to erroneous modeling assumptions.

This thesis is organized as follows. In Chapter 2, we develop a mathematical representation of covariation analysis in terms of composition and scale. In Chapter 3, we use this representation to explore how error in the unmeasured scale parts propagate into correlation estimates. Finally, in Chapter 4, we discuss the potential future directions for this work and provide guidance to researchers looking to perform correlation analyses using high-throughput genomic data.

Chapter 2
Covariation Analysis in Terms of Scale

2.1 Mathematical Representation of Covariation Analysis

In this section, we will derive a mathematical representation of covariation analysis in terms of composition and scale based on Nixon's work [9]. In chapter 1, we have defined scale ($W_n^{\perp}$) and composition of d-th taxa ($W_{dn}^{\parallel}$) in the community n. Let $W_{d_1\cdot}^{\parallel} = (W_{d_1 1}^{\parallel}, \ldots, W_{d_1 N}^{\parallel})$ denote a vector which contains the composition information of d_1-th taxa in all N communities, and $W^{\perp} = (W_1^{\perp}, \ldots, W_N^{\parallel})$ is also a vector that contains the scale information of all N communities. In order to explore how error in unmeasured scale contribute to our correlation estimates, first we need to break equation 1.1 into scale and compositional parts. To simplify the notation, let $\hat{W}$ be a shorhand for $\log W$.

$$\begin{aligned}
\Sigma_{d_1 d_2} &= Cov(\log(W_{d_1\cdot}^{\parallel} \cdot W^{\perp}), \log(W_{d_2\cdot}^{\parallel} \cdot W^{\perp})) \\
&= Cov(\hat{W}_{d_1\cdot}^{\parallel} + \hat{W}^{\perp}, \hat{W}_{d_2\cdot}^{\parallel} + \hat{W}^{\perp}) \\
&= \sigma_{\hat{W}_{d_1}^{\parallel}} \sigma_{\hat{W}_{d_2}^{\parallel}} \rho(\hat{W}_{d_1\cdot}^{\parallel}, \hat{W}_{d_2\cdot}^{\parallel}) + \sigma_{\hat{W}_{d_1}^{\parallel}} \sigma_{\hat{W}^{\perp}} \rho(\hat{W}_{d_1\cdot}^{\parallel}, \hat{W}^{\perp}) \\
&+ \sigma_{\hat{W}_{d_2}^{\parallel}} \sigma_{\hat{W}^{\perp}} \rho(\hat{W}_{d_2\cdot}^{\parallel}, \hat{W}^{\perp}) + \sigma_{\hat{W}^{\perp}}^2. \qquad (2.1)
\end{aligned}$$

Above we have expressed covariation analysis in terms of the scale and compositional parts. For compositional data, we have little knowledge of scale information ($W^{\perp}$) from the measurements, hence there are three unknowns, $\sigma_{\hat{W}^{\perp}}$, $\rho(\hat{W}_{d_1\cdot}^{\parallel}, \hat{W}^{\perp})$ and $\rho(\hat{W}_{d_2\cdot}^{\parallel}, \hat{W}^{\perp})$ that stop us from estimating target correlation. In this case, the uncertainty in scale parts may make our final estimand unreliable and it is of significance to understand what effect the lack of scale information will have on our final estimand. So we will explore

how the error in these three unknowns propagate into target estimand.

2.2 Covariation Analysis and Scale Information

Since $\sigma_{\hat{W}^\perp}$, $\rho(\hat{W}^\parallel_{d_1\cdot}, \hat{W}^\perp)$ and $\rho(\hat{W}^\parallel_{d_2\cdot}, \hat{W}^\perp)$ are necessary to estimate correlation, an estimator must make an assumption of these variables directly or indirectly. However, some estimators are black-box which do not mention how they assume these variables. Thus we develop a method for researchers to trace back and find the assumption an estimator makes. So the question will be: given an estimator and the correlation it estimates, can we uniquely solve for the variables above? However, it is impossible to solve for $\sigma_{\hat{W}^\perp}$ if we fix $\rho(\hat{W}^\parallel_{d_1\cdot}, \hat{W}^\perp)$ and $\rho(\hat{W}^\parallel_{d_2\cdot}, \hat{W}^\perp)$ in equation 2.1 for it is a quadratic equation of $\sigma_{\hat{W}^\perp}$. Similarly, we cannot solve for $\rho(\hat{W}^\parallel_{d_1\cdot}, \hat{W}^\perp)$ and $\rho(\hat{W}^\parallel_{d_2\cdot}, \hat{W}^\perp)$ at the same time given $\sigma_{\hat{W}^\perp}$ since there are two unknowns. Thus we need to address this problem from a more general perspective, and rewrite equation 2.1 into a vector form:

$$\Sigma = \Gamma + \sigma_{\hat{W}^\perp}(\sigma_{\hat{W}^\parallel} \circ \rho(\hat{W}^\perp, \hat{W}^\parallel))\mathbf{1}^T + \sigma_{\hat{W}^\perp}\mathbf{1}(\sigma_{\hat{W}^\parallel} \circ \rho(\hat{W}^\perp, \hat{W}^\parallel))^T + \sigma^2_{\hat{W}^\perp}\mathbf{1}_{D\times D}. \quad (2.2)$$

where $\sigma_{\hat{W}^\perp}$ is a scalar. $\hat{W}^\parallel$ and $\rho(\hat{W}^\parallel, \hat{W}^\perp)$ are $D \times 1$ vectors. Γ is a $D \times D$ matrix and $\Gamma = Cov(\hat{W}^\parallel, \hat{W}^\parallel)$. $\circ$ is Hadamard product. $\mathbf{1}$ is a $D \times 1$ vector. $\mathbf{1}_{D\times D}$ is a $D \times D$ matrix. To simplify the notation, let $\alpha = \sigma_{\hat{w}^\parallel} \circ \rho(\hat{W}^\perp, \hat{W}^\parallel)$, and $a = \sigma_{\hat{w}^\perp}$. Then rearrange the equation 2.2, we have.

$$\Sigma - \Gamma = a(\alpha + a\mathbf{1})\mathbf{1}^T + a\mathbf{1}\alpha^T. \quad (2.3)$$

Consider the case $a \neq 0$, and multiply $\mathbf{1}$ from the right side.

$$\begin{aligned}(\Sigma - \Gamma)\mathbf{1} &= a(\alpha + a\mathbf{1})\mathbf{1}^T\mathbf{1} + a\mathbf{1}\alpha^T\mathbf{1} \\ &= aD(\alpha + a\mathbf{1}) + aM\mathbf{1} \quad (2.4)\end{aligned}$$

Here D is the dimension of Σ (i.e., the number of taxa). Let $M = \sum_d^D \alpha_d$. From equation 2.4, if a is known and α is unknown, our aim is to solve for α:

$$\alpha = \frac{1}{aD}(\Sigma - \Gamma)\mathbf{1} - \frac{M}{D}\mathbf{1} - a\mathbf{1}. \quad (2.5)$$

Then multiply $\mathbf{1}^T$ from left side, we have

$$M = \frac{1}{aD}\mathbf{1}^T(\Sigma - \Gamma)\mathbf{1} - M - aD,$$
$$M = \frac{1}{2aD}\mathbf{1}^T(\Sigma - \Gamma)\mathbf{1} - \frac{aD}{2}. \tag{2.6}$$

Plug equation 2.5 into equation 2.6 and let $S = \sum_{d_1}^{D}\sum_{d_2}^{D}(\Sigma_{d_1 d_2} - \Gamma_{d_1 d_2}) = \mathbf{1}^T(\Sigma - \Gamma)\mathbf{1}$, we have a unique solution for α:

$$\alpha = \frac{1}{aD}(\Sigma - \Gamma)\mathbf{1} - \frac{S}{2aD^2}\mathbf{1} + \frac{a}{2}\mathbf{1}. \tag{2.7}$$

Thus we have proved if $\sigma_{\hat{w}^\perp}$ is fixed, $\rho(\hat{W}^\perp, \hat{W}^\parallel))$ and the final estimate Σ is a one-to-one mapping. Next let us explore the relation between Σ and $\sigma_{\hat{w}^\perp}$. Consider the case where α is known and a is unknown. Rearrange equation 2.4, we have

$$D^2a^2 + 2DMa - S = 0.$$

Thus we can get

$$a = \frac{-M \pm \sqrt{M^2 + S}}{D}.$$

Since a must be non-negative, there is a unique solution of a

$$a = \frac{-M + \sqrt{M^2 + S}}{D}. \tag{2.8}$$

Overall, we have proved that for a specific Σ, there is always a unique solution to $\sigma_{\hat{W}^\perp}$ or $\rho(\hat{W}^\parallel, \hat{W}^\perp)$ if we know one of them. Also, if we have knowledge of M, we can also uniquely solve for $\sigma_{\hat{W}^\perp}$ and $\rho(\hat{W}^\parallel, \hat{W}^\perp)$. This indicates that if we know the assumption an estimator makes (such as mean$(M) = 0$), then we can trace back and figure out what the assumption they make of $\sigma_{\hat{W}^\perp}$ and $\rho(\hat{W}^\parallel, \hat{W}^\perp)$. Furthermore, we can see how the error in these variables related to scale influence our correlation estimates. Therefore, in the next chapter we will do a sensitivity analysis and show how the error in $\sigma_{\hat{W}^\perp}$ propagates into the correlation estimates.

Chapter 3
Sensitivity Analysis of Correlation Estimation

3.1 Error in Scale Information

In the previous chapter, we have proved it is applicable to conduct the sensitivity analysis of correlation estimation. In this section, we will demonstrate how error in $\sigma_{\hat{W}^\perp}$ propagates into our target estimand $\Sigma_{d_1 d_2}$. As what we do to simplify the notation in chapter 2, let $\alpha = \sigma_{\hat{w}^\parallel} \circ \rho(\hat{W}^\perp, \hat{W}^\parallel))$, and $a = \sigma_{\hat{w}^\perp}$.

$$\Sigma_{d_1 d_2} = \Gamma_{d_1 d_2} + a^2 + a(\alpha_{d_1} + \alpha_{d_2}). \tag{3.1}$$

Ideally, if we have knowledge of the exact scale of the system, then $\sigma_{\hat{w}^\perp}$ can be calculated. However, that is not possible since the real abundance is unavailable. Therefore, the real $\sigma_{\hat{w}^\perp}$ is hard to reach and if we make an assumption of $\sigma_{\hat{w}^\perp}$ there will be an error in it. Let ϵ denote the difference between our assumption and the true $\sigma_{\hat{w}^\perp}$. a is the true value of $\sigma_{\hat{w}^\perp}$, thus the disturbed standard deviation will be $a + \epsilon$. Since standard deviation is always non-negative, ϵ should take values in $[-a, \infty)$. Let Σ^+ be the disturbed esitimand:

$$\Sigma^+_{d_1 d_2} = \Gamma_{d_1 d_2} + (a + \epsilon)^2 + (a + \epsilon)(\alpha_{d_2} + \alpha_{d_2}). \tag{3.2}$$

Then the error in estimand will be:

$$\Sigma_{d_1 d_2} - \Sigma^+_{d_1 d_2} = -(2a\epsilon + \epsilon^2) - \epsilon(\alpha_{d_1} + \alpha_{d_2}). \tag{3.3}$$

Obviously, the error in $\Sigma_{d_1d_2}$ is a quadratic function of ϵ. To find the upper bound of $\Sigma_{d_1d_2}$, rearrange the equation,

$$\Sigma_{d_1d_2} = \Sigma^+_{d_1d_2} - \left[\epsilon + (a + \frac{\alpha_{d_1} + \alpha_{d_2}}{2})\right]^2 + (a + \frac{\alpha_{d_1} + \alpha_{d_2}}{2})^2. \tag{3.4}$$

So $\Sigma_{d_1d_2}$ can be considered as a function of $\Sigma^+_{d_1d_2}$ and ϵ, i.e. $f(\Sigma^+_{d_1d_2}, \epsilon)$. Since $\epsilon \in [-a, \infty)$, $\Sigma_{d_1d_2}$ is a parabola when $\alpha_{d_1} + \alpha_{d_2} > 0$ and takes values in $(-\infty, (a + \frac{\alpha_{d_1}+\alpha_{d_2}}{2})^2 + \Sigma^+_{d_1d_2}]$. It reaches its maximum when $\epsilon = -(a + \frac{\alpha_{d_1}+\alpha_{d_2}}{2})$. At the maximum, plug $\epsilon = -(a + \frac{\alpha_{d_1}+\alpha_{d_2}}{2})$ into equation 3.2 and all the terms containing a will vanish which indicates the disturbed estimand $\Sigma^+_{d_1d_2}$ is independent of $\sigma_{\hat{w}^\perp}$. Furthermore, if $\alpha_{d_1} + \alpha_{d_2} < -2a$, as the red line in figure 3.1 shows, it is possible that the estimand is the same as the real one even though there exists error in $\sigma_{\hat{w}^\perp}$. Instead, if $\alpha_{d_1} + \alpha_{d_2} \leq 0$, the error in final estimand $\Sigma_{d_1d_2}$ will be greater with the increase of the error in $\sigma_{\hat{w}^\perp}$.

To verify our conclusion, we conduct the following simulation. First, we generate a covariance matrix Σ. Second, we generate compositional information with the covariance matrix Σ, and sample with the compositional information. The detailed algorithm is as follows:

Algorithm 1 Sample Generation

function Sample Generation

$D \leftarrow 4$, # Set the dimension of covariance matrix as 4.

$\Sigma \leftarrow IW(D + 3, I_D)$, # Generate covariance matrix using Inverse-Wishart distribution.

$\mu \leftarrow N(0, 4I_D)$,. # Generate mean value of samples with normal distribution.

$\eta \leftarrow N(\mu, \Sigma)$,

$w_{ij} \leftarrow \exp(\eta_{ij})$, # Generate non-negative samples using multivariate-normal distribution.

$w_i^{\|} \leftarrow \frac{w_{ij}}{\sum_j w_{ij}}$,# Generate compositional information.

$Y_i \leftarrow Mult(w_i^{\|}, n = 1e5, D = 4)$, # Sample with the compositional information adopting multinomial distribution.

$\hat{w_i}^{\|} \leftarrow \frac{Y_{ij}}{\sum_j Y_{ij}}$.# Calculate the compositional information from samples.

$\Gamma = Cov(\log(\hat{w}^{\|}))$.

end function

To find out how errors propagate, we make the following assumption: $M = \sum_d^D \alpha_d = 0$. Then we adopt ccLasso [4] to estimate $\Sigma_{d_1d_2}$, and calculate α and a based on equation 2.5 and 2.7. Afterwards, we switch the value of ϵ and add it to a. Finally, we calculate the disturbed covariation matrix Σ^+ based on equation 3.2 and see the difference between

undisturbed correlation estimates $\Sigma_{d_1 d_2}$ and the disturbed one $\Sigma^{+}_{d_1 d_2}$.

Algorithm 2 Sensitivity Analysis

function SENSITIVITY ANALYSIS($\hat{w}^{\parallel}, \Gamma, D$)
$M \leftarrow 0$,
$\Sigma \leftarrow ccLasso(\hat{w}^{\parallel})$, # Undisturbed correlation estimates.
$S \leftarrow sum(Sigma - Gamma)$,
$a \leftarrow \sqrt{S}/D$,. # Calculate a.
$\alpha \leftarrow rowSums(\Sigma - \Gamma)/(a * D) - a * \mathbf{1}$, #Calculate α.
$\epsilon \leftarrow [-a, 1]$, # Change value of ϵ.
$\Sigma^{+}_{d_1 d_2} \leftarrow \Gamma_{d_1 d_2} + (a+\epsilon)^2 + (a+\epsilon)(\alpha_{d_1} + \alpha_{d_2})$,# Calculate disturbed correlation estimates.
Calculate the difference between $\Sigma_{d_1 d_2}$ and $\Sigma^{+}_{d_1 d_2}$.
end function

The results are as Figure 3.1 which demonstrate our conclusion above.

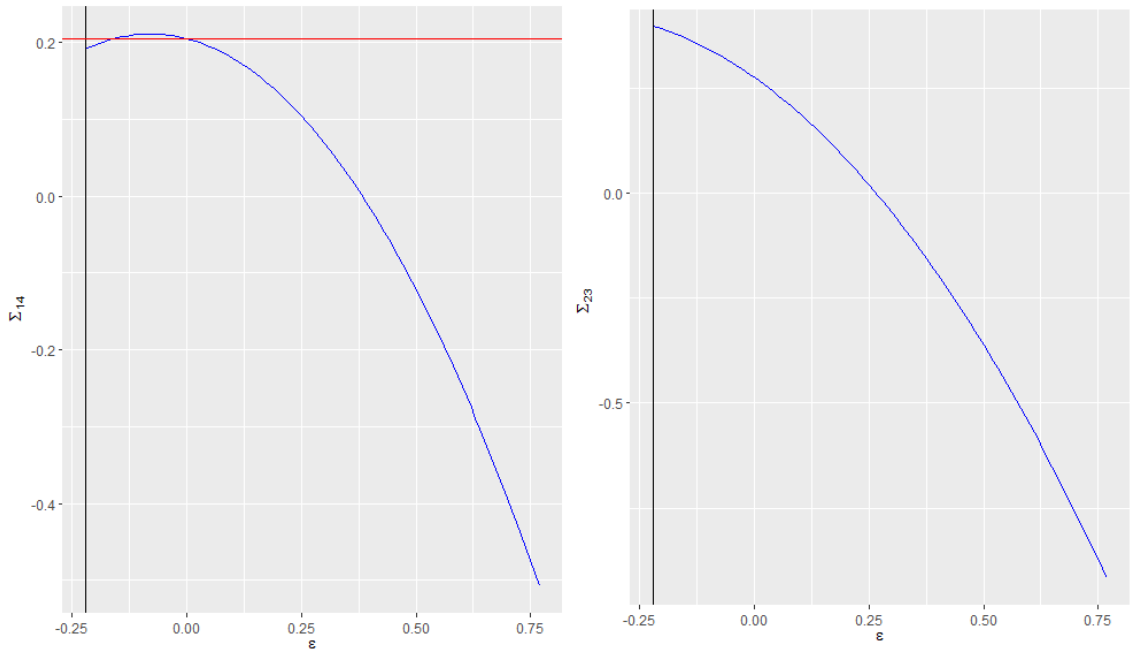

Figure 3.1. x-aixs is the error ϵ, y-axis is the estimated correlation Σ_{14} where $\alpha_1 + \alpha_4 > 0$ (left), and Σ_{23} where $\alpha_2 + \alpha_3 \leq 0$ (right). Black line is the minimum value of ϵ to ensure $\sigma_{\hat{w}^{\perp}} + \epsilon \geq 0$. Red line is the value of Σ_{14} when ϵ takes value of 0.

Overall, we have proved that the error in correlation estimates $\Sigma_{d_1 d_2}$ can be regarded as a quadratic function of ϵ. In this case, since some techniques first make an assumption of $\sigma_{\hat{w}^{\perp}}$ in order to estimate the target correlation, if the error in assumed $\sigma_{\hat{w}^{\perp}}$ is huge, the final estimates will be unreliable according to our result above. This warns us that the

erroneous assumption of $\sigma_{\hat{w}^\perp}$ will be disastrous to the final correlation estimates. And a researcher can adopt our method to evaluate whether an estimator will be applicable to their task by judging how much the assumption differs from the reality based on their prior knowledge.

Besides, we show that when $\alpha_{d_1} + \alpha_{d_2} > 0$ and $\epsilon = -(a + \frac{\alpha_{d_1}+\alpha_{d_2}}{2})$, the disturbed estimand $\Sigma^+_{d_1d_2}$ will be independent of $\sigma_{\hat{w}^\perp}$.

$$\Sigma^+_{d_1d_2} = \Gamma_{d_1d_2} + \frac{3}{4}(\alpha_{d_1} + \alpha_{d_2})^2. \tag{3.5}$$

This finding has a possible application. We demonstrate the error in estimand will be huge if there is a lot of uncertainty in $\sigma_{\hat{w}^\perp}$,i.e. $|\epsilon| >> \sigma_{\hat{w}^\perp}$. To address this problem, we can make a conservative estimation $\Sigma^+_{d_1d_2} = \Gamma_{d_1d_2} + \frac{3}{4}(\alpha_{d_1} + \alpha_{d_2})^2$ with a prior knowledge of α_{d_1} and α_{d_2} for this estimation does not require knowledge of $\sigma_{\hat{w}^\perp}$. Though there is still error $\Sigma_{d_1d_2} - \Sigma^+_{d_1d_2} = (a + \frac{\alpha_{d_1}+\alpha_{d_2}}{2})^2$ in the final estimand, it will be better than the result when $\epsilon >> \sigma_{\hat{w}^\perp}$ and may be acceptable under certain situations.

3.2 Lower bound of Correlation Estimation and Its Application

We have proved the disturbed estimand Σ^+ is a quadratic function of ϵ, thus there will be a lower bound for it. Therefore, if we can ensure this lower bound positive, the estimand will always be positive regardless of the value of error in $\sigma_{\hat{w}^\perp}$. In this section, we will explore the lower bound of Σ^+ and how it can be applied in practice.

$$\Sigma^+_{d_1d_2} = \Gamma_{d_1d_2} + \left[(a+\epsilon) + \frac{\alpha_{d_1} + \alpha_{d_2}}{2}\right]^2 - \left(\frac{\alpha_{d_1} + \alpha_{d_2}}{2}\right)^2 \tag{3.6}$$

If $\Gamma_{d_1d_2} > \left(\frac{\alpha_{d_1}+\alpha_{d_2}}{2}\right)^2$ is ensured, no matter what value the error ϵ takes, the correlation between taxa d_1 and d_2 is always estimated to be positive. In practice, we commonly have access to $\sigma_{\hat{w}^\parallel}$. Thus if $\Gamma_{d_1d_2} > \left(\frac{|\sigma_{\hat{w}^\parallel_{d_1}}| + |\sigma_{\hat{w}^\parallel_{d_2}}|}{2}\right)^2$ is guaranteed, then we can always make a positive estimation of $\Sigma_{d_1d_2}$ for $-1 \le \rho \le 1$. We shall mention that this situation only happens in a rare case when $\Sigma_{d_1d_2} = \Sigma_{d_1d_1} = \Sigma_{d_2d_2}$. This result indicates that for a very strong positive correlation, we can always make an accurate estimation no matter what the error in scale information is.

Chapter 4
Possible Application and Future Direction

4.1 Discussion

In the previous chapter, we have demonstrated how our framework make it possible to perform sensitivity analysis. Besides, there also exist some interesting subareas that can be explored, and in this section we will give an example to show readers some possible future directions.

Think back to the Section 2.3 where we introduce a variable $M = \sum_d^D (\sigma_{\hat{W}_d^{\parallel}} \circ \rho(\hat{W}_{d\cdot}^{\parallel}, \hat{W}^{\perp}))$, thus $M = \frac{\sum_d^D Cov(\hat{W}_{d\cdot}^{\parallel}, \hat{W}^{\perp})}{D \cdot \sigma_{\hat{W}^{\perp}}}$. In Chapter 3, we make an assumption that $M = 0$ which indicates that the mean of $\frac{Cov(\hat{W}_{d\cdot}^{\parallel}, \hat{W}^{\perp})}{\sigma_{\hat{W}^{\perp}}}$ equals to zero. This assumption makes sense in the context where the mean of $Cov(\hat{W}_{d\cdot}^{\parallel}, \hat{W}^{\perp})$ is around zero or $Cov(\hat{W}_{d\cdot}^{\parallel}, \hat{W}^{\perp})$ is relatively small compared to $\sigma_{\hat{W}^{\perp}}$. From equation 2.8

$$\sigma_{\hat{W}^{\perp}} = \frac{-M + \sqrt{M^2 + S}}{D}.$$

where $S = \sum_{d_1}^D \sum_{d_2}^D (\Sigma_{d_1 d_2} - \Gamma_{d_1 d_2}) = \mathbf{1}^T(\Sigma - \Gamma)\mathbf{1}$. S can be regarded as a function of M and $\sigma_{\hat{W}^{\perp}}$. If we have some prior knowledge of M and $\sigma_{\hat{W}^{\perp}}$, then the mean correlation i.e. $Mean(\Sigma_{d_1 d_2})$ will be available. This method lowers the difficulty of predicting the mean correlation in Σ since we only need to have knowledge of two scalars instead of learning a matrix with high dimensions.

Although we have demonstrated how error in scale information propagates to the final estimand, the assumption we make, i.e., $M = 0$ may not be applicable to all the situations. Another possible further work might be represent estimand as a posterior

distribution of M and explore how error in $\sigma_{\hat{W}^{\perp}}$ propagates under this circumstance. Besides, how error in $\rho(\hat{W}^{\parallel}, \hat{W}^{\perp})$ affects the final result also needs to explored.

4.2 Conclusion

In this work, we first form a mathematical representation of convariation analysis in terms of scale and composition. Next, we show a method to find out what assumption an estimator makes of $\sigma_{\hat{W}^{\perp}}$ and $\rho(\hat{W}^{\parallel}, \hat{W}^{\perp})$. Then we show how the error in unmeasured scale parts can propagate to correlation estimates, which will help researchers determine the applicability of an estimator. At last, we discuss about the possible future improvement our work.

Bibliography

[1] Aitchison,J. and Shen,S.M. (1980). "Logistic-normal distributions: Some properties and uses". *Biometrika*, 67:261–272.

[2] Aitchison, J. (2003) "The statistical analysis of compositional data". Caldwell, New Jersey, USA: Blackburn Press, 416 pp.

[3] Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (2006) "Compositional data analysis in the geosciences : from theory to practice". Number 24 in Special Publication. London, UK: Geological Society. 224 pp.

[4] Fang, H., Deng, M., and et al. (2015). "CCLasso: correlation inference for compositional data through Lasso". *Bioinformatics.*, 31(19):3172-3180.

[5] Friedman, J. and Alm, E. J. (2012). "Inferring correlation networks from genomic survey data". *PLoS Computational Biology*, 8(9):e1002687.

[6] Jackson D (1997) "Compositional data in community ecology: the paradigm or peril of proportions?" Ecology 78: 929–940.

[7] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., Bonneau, R. A. (2015) "Sparse and Compositionally Robust Inference of Microbial Ecological Networks". *PLoS Computational Biology* 11(5): e1004226.

[8] Medini, D., Serruto, D., Parkhill, J., Relman, D., Donati, C. and et al. (2008) "Microbiology in the post-genomic era". *Nat Rev Microbiol*, 6:419–430.

[9] Nixon, M. P., Letourneau, J., David, L. A., Lazar, N. A., Mukherjee, S., and Silverman, J.D. (2022) Scale Reliant Inference". *arXiv.* https://doi.org/10.48550/arXiv.2201.03616.

[10] Pawlowsky-Glahn, V., Buccianti, A. (2011) "Compositional data analysis : theory and applications". Chichester, West Sussex, UK: Wiley. 400 pp.

[11] Quinn, T. P., Richardson, M. F., Lovell, D. et al. (2017) "propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis". *Sci Rep* **7**, 16252 .

[12] Schwager, E., Mallick, H., Ventz, S., Huttenhower, C. (2017) "A Bayesian method for detecting pairwise associations in compositional data". *PLoS Computational Biology*, 13(11):

[13] Silverman, J. D., Bloom, R. J., Jiang, S., Durand, H. K., Dallow, E., Mukherjee, S., and David, L. A. (2021). "Measuring and mitigating pcr bias in microbiota datasets". *PLoS Computational Biology*, 17(7):e1009113.