UCLouvain

AGRO LOUVAIN

Faculté des bioingénieurs

Inferring cohort differences in microbiota composition using Latent Dirichlet allocation model

Author : Maxime CORNEZ
Supervisor : Prof. Laura SYMUL
Readers : PhD Manon MARTIN
PhD Jérôme AMBROISE

Academic year : 2025 - 2026

Master's Thesis submitted in fulfillment of the requirements for the Degree of Bioscience engineering: Master in Agricultural Sciences

Acknowledgments

First and foremost, I would like to thank Professor Laura Symul for giving me the opportunity to conduct my Master's thesis under her supervision. It has been an enriching experience on every level of my life as a young researcher. I thank her for all the knowledge she has imparted to me, for her invaluable advice throughout this thesis, and especially for her patience and the reassurance she provided whenever I began to doubt myself.

I would also like to thank Dr. Philippe Hauchamps. Our invaluable discussions often brought to light my own blind spots on the subject, and you guided me toward aspects of my thesis I had not initially considered. Our exchanges were immensely precious to me, and I hope you will take the continuation of this project as far as you desire.

I also wish to thank everyone in the "Working with Laura" (I love the name) group for welcoming me during the realization of my thesis. Each presentation was an opportunity for me to learn more about a subject that has fascinated me for a long time.

My thanks also go to my readers, Dr. Manon Martin and Dr. Jérôme Ambroise. I am deeply grateful to you for finding the time in your busy schedules to read my work and for participating in the completion of this major milestone in my studies.

I would like to thank all the professors, teaching assistants, and staff members of the Faculty of Bioscience Engineering at UCLouvain. Within this faculty, I was able to flourish as a student and an apprentice researcher; I learned a tremendous amount of things (even if I did not always retain them all). This faculty is filled with excellent professors and assistants who, through their benevolent guidance, constantly pushed me to go further. This is truly invaluable.

On a more personal note, I want to thank my dad. We may be far apart geographically, but I have always been able to count on him whenever I needed his presence. We are a team!

I also want to write a special word for my mom. Even though she is no longer here to see me complete this thesis, I wanted to dedicate a part of this work to her.

To my two companions in adventure, Loïc and Raphaël: without you, I surely would not have reached the end of this journey, not like that at least. Thank you for the lunch breaks, the often-too-long coffee breaks, the often-too-short evenings spent together, and all the other moments you were there for me. I could not dream of better friends.

And finally, to my fiancée Chloé: I think there were times when you believed I would finish this thesis more than I did myself. Your unwavering faith in me has been my greatest driving force every single day. Thank you for being you, and thank you for being by my side.

Declaration of Contributions (CRediT Taxonomy)

	Maxime Cornez	**Other contributors**	**Artificial Intelligence**
Conceptualisation	X	Prof. Laura Symul	
Methodology	X	Prof. Laura Symul and PhD. Philippe Hauchamps	
Investigation	X	Prof. Laura Symul	Research Rabbit
Formal analysis	X	Prof. Laura Symul	Google Gemini : Assistance with troubleshooting computer issues
Validation	X	Prof. Laura Symul and PhD. Philippe Hauchamps	
Editorial – original version	X		DeepL
Writing – Proofreading and Editing	X	Prof. Laura Symul	Google Gemini : Rewriting, syntactic corrections, translation, and adjustments to the LaTeX typesetting code
Visualization	X		
Supervision		Prof. Laura Symul and PhD. Philippe Hauchamps	

Detailed description of contributions:

- **Conceptualization:** Formulation of the research scope, objectives, and overall feasibility through collaborative discussions between the author and Prof. Laura Symul.
- **Methodology:** Exploration and selection of the most appropriate analytical approaches through discussions with Prof. L. Symul and PhD Philippe Hauchamps. Prof. L. Symul provided initial guidance, followed by independent methodological research by the author and subsequent validation.
- **Investigation:** Conduction of the literature review, building upon foundational sources initially provided by Prof. L. Symul. The AI-powered tool *Research Rabbit* was then used by the author to discover further relevant literature, with rigorous manual screening and verification of every suggested source. *Zotero* was used for reference management.
- **Formal analysis:** Execution of the statistical analyses and development of the analytical code. Analytical strategies were iteratively discussed with Prof. L. Symul, who provided critical feedback on the author's proposals and suggested alternative methods when analytical challenges arose. Generative AI (*Google Gemini*) was utilized strictly for debugging and syntax correction when encountering technical blocks. All AI-proposed code was systematically reviewed, tested, and validated by the author, who assumes full responsibility for the final implementation.
- **Validation:** Continuous assessment and validation of the analytical results during regular supervisory meetings with Prof. L. Symul and PhD P. Hauchamps.
- **Writing – original draft:** Independent authoring of the initial manuscript. *DeepL* and online searches were used strictly for targeted English vocabulary translation.
- **Writing – review and editing:** Incorporation of extensive structural and content revisions based on feedback from Prof. L. Symul (including text improvements, content additions, and graph formatting). Generative AI (*Google Gemini*) was used for LaTeX formatting assistance and occasional sentence restructuring. All AI-assisted phrasing was systematically reviewed, manually edited, and validated by the author, who assumes full responsibility for the entirety of the written content.
- **Visualization:** Independent development of the data visualization code, supported by online technical forums (e.g., *StackOverflow*). Generative AI (*Google Gemini*) was utilized strictly to resolve specific graphical formatting details. All AI-proposed code was systematically reviewed and validated by the author, who assumes full responsibility for the generated figures.
- **Supervision:** Prof. L. Symul acted as the primary thesis promoter, overseeing the research planning and overall direction. PhD P. Hauchamps acted as a scientific advisor, providing specialized expertise and continuous guidance on specific topics.

Abstract

Microbiota composition data are inherently characterized by high sparsity and stochastic overdispersion, complicating the reliable detection of cohort-level differences. To address this challenge, this thesis evaluates the capacity of a joint analytical framework, combining Latent Dirichlet Allocation (LDA) and PERMANOVA, to infer and test global differences between microbiota cohorts. An *in silico* generative process coupled with a full factorial design is implemented to evaluate the pipeline against controlled variations in biological signal amplitude (A), cohort size (N_g), and the Negative Binomial distribution size parameter (ϕ). By exclusively simulating the cohort effect on the latent topic prevalence matrix (Γ), this framework isolates the statistical behavior of the pipeline from the confounding factors of empirical datasets.

The systematic evaluation reveals a crucial functional decoupling between latent space reconstruction and statistical testing. First, the structural fidelity of the LDA inference is primarily dictated by overdispersion; extreme variance ($\phi \leq 1$) degrades the accuracy of the inferred matrices, a degradation that increased cohort recruitment alone cannot fully compensate for. Despite this, the downstream statistical testing phase demonstrates consistent reliability. The framework maintains a controlled Type I error rate (oscillating around 5%), ensuring that extreme overdispersion does not translate into spurious cohort separations. Furthermore, the statistical power is primarily governed by the biological effect amplitude and cohort size, exhibiting compensatory dynamics that facilitate accurate detection provided the biological contrast is pronounced or sampling is sufficient. Consequently, this thesis quantitatively maps the operational limits of the LDA-PERMANOVA pipeline, providing data-driven guidelines for the experimental design of future microbiota studies.

Contents

List of Figures

List of Tables

List of Abbreviations

A Amplitude.

N_g Number of samples per group.

ϕ Size parameter.

ANOSIM Analysis Of Similarities.

ANOVA Analysis Of Variance.

ASV Amplicon Sequence Variant.

CST Community Sub Types.

DESeq Differential Expression analysis for Sequencing.

Df Degrees of freedom.

DMM Dirichlet Multinomial Mixture.

DMR Dirichlet Multinomial Regression.

DTM Document-Term Matrix.

FDR False Discovery Rate.

FPR False Positive Rate.

GaP Gamma-Poisson.

GLM Generalized Linear Model.

JSD Jensen-Shannon Divergence.

LDA Latent Dirichlet Allocation.

LDAcov Latent Dirichlet Allocation with covariates.

LR Likelihood Ratio.

MANOVA Multivariate Analysis Of Variance.

MCMC Markov Chain Monte Carlo.

MULT Multinomial.

NB Negative Binomial.

NLP Natural Language Processing.

NMF Non-negative Matrix Factorization.

OR Odd Ratio.

OTU Operational Taxon Unit.

PAM Partitioning Around Medoids.

PCA Principal Composant Analysis.

PCoA Principal Coordinate Analysis.

PCR Polymerase Chain Reaction.

PERMANOVA Permutational Multivariate Analysis of Variance.

RMSE Root Mean Square Error.

STM Structural Topics Model.

TMM Trimmed Mean of M-values.

VEM Variational Expectation-Maximization.

ZINB Zero-Inflated Negative Binomial.

1. Introduction

1.1 The Human Microbiota: Ecological Foundations and Health Implications

Microorganisms, particularly bacteria, are omnipresent across terrestrial ecosystems, displaying significant abundance and diversity The understanding of this microscopic world has continuously evolved since the initial observations by Antonie van Leeuwenhoek in the 17th century. While historical research initially focused on the pathogenic nature of microbes, notably through the foundational work of Robert Koch and Louis Pasteur, the scientific perspective has progressively broadened. It is now established that bacteria fulfill diverse and essential functional roles that extend far beyond disease transmission (Berche 2007; Berg et al. 2020; Opal 2010).

This conceptual evolution led, in 2001, to the first formal definition of the term microbiota by Lederberg and McCray (2001). This concept denotes the specific ecological community of commensal, symbiotic, and pathogenic microorganisms that share a defined environmental space.

The relevance of these communities is particularly evident within the human body. Recent quantitative estimates illustrate this extensive presence: the human body appears to harbor slightly more bacterial cells than actual human cells, with an estimated ratio of approximately 40×10^{12} bacterial cells to 30×10^{12} human cells (Sender et al. 2016).

This paradigm shift led Lederberg and McCray (2001) to formally define the *microbiome* as the specific ecological community of commensal, symbiotic, and pathogenic microorganisms sharing a defined environmental space. Definitions have since evolved and proliferated (Berg et al. 2020). For the purposes of this study, the **microbiome** is defined as encompassing both the microbial community and its entire functional environment, including all of its genomes (Berg et al. 2020; Liu 2016). Within this broader ecosystem, the specific collection of living microorganisms is called the **microbiota**.

The significance of these microbial populations is highly pronounced within the human body. Recent quantitative estimates indicate that the human host harbors slightly more bacterial cells than human cells, with a ratio of approximately 40×10^{12} to 30×10^{12} (Sender et al. 2016).

Consequently, the human host is conceptualized as a "supra-organism" or holobiont, sustained by a network of coexisting microbial communities (Turnbaugh et al. 2007). These localized ecosystems, including the gut, skin, and vaginal microbiota (Figure 1.1), interact dynamically with the host and with one another. Despite these interactions, each anatomical niche maintains a distinct composition and functional profile shaped by its local environment (Hou et al. 2022). Furthermore, these resident bacteria exert essential physiological effects, driving processes such as the degradation of complex compounds, the maturation of the immune system, and metabolic regulation (Belkaid and Harrison

2017; Fan and Pedersen 2021).

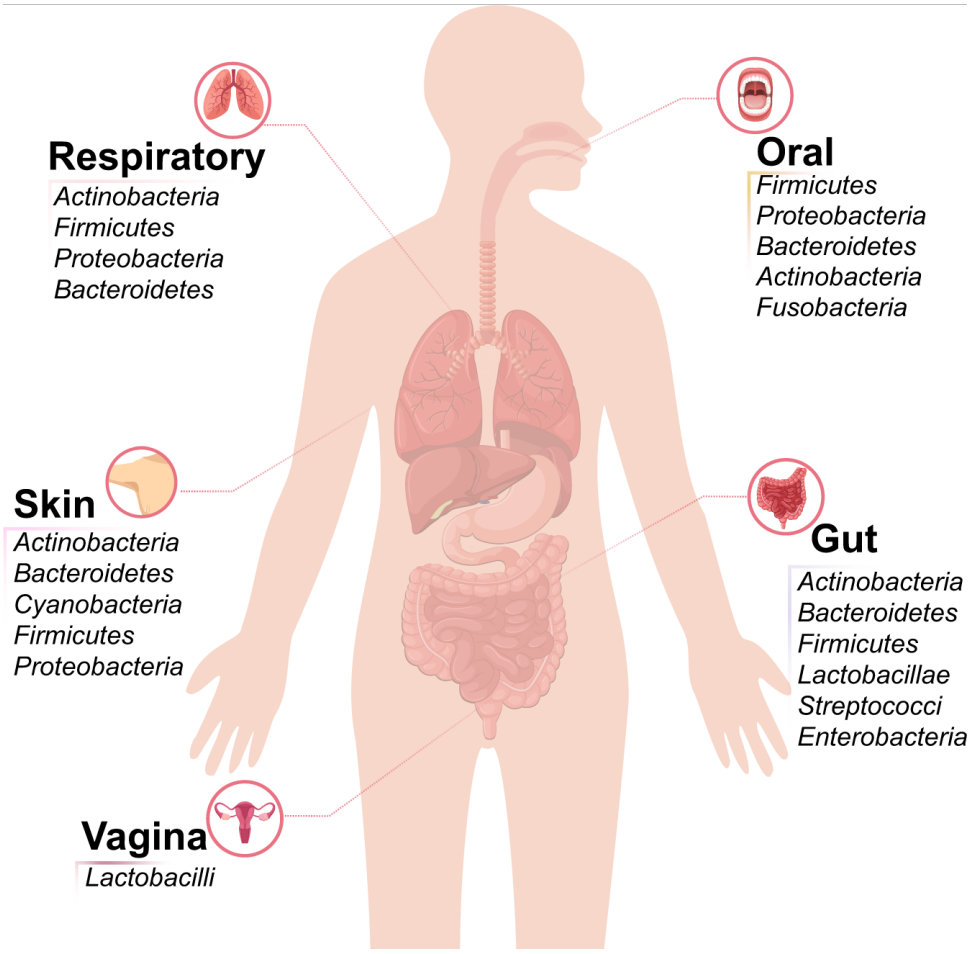


Figure 1.1: **Human microbiota composition across different anatomical niches.** The predominant bacterial taxa are specific to their local environment, such as the oral cavity, respiratory tract, skin, gut, and vagina. Reproduced from Hou et al. (2022).

Since optimal host health relies on this symbiosis, the disruption of this microbial equilibrium, termed *dysbiosis*, can lead to adverse health outcomes. In the gut, for instance, dysbiosis is associated with chronic inflammatory conditions, such as Crohn's disease (Manichanh et al. 2012), and obesity (Fan and Pedersen 2021). Within the vaginal microbiota, such imbalances are associated with severe obstetric complications, including preterm birth and miscarriage, as well as an increased host vulnerability to infections, notably HIV (Symul et al. 2023).

In light of these health implications, the study of the human microbiota has become a crucial area of research. High-throughput sequencing now enable the capture of its diversity, generating substantial volumes of data. However, analyzing these datasets involves navigating a dual biological and statistical reality. Biologically, even subtle variations in the microbial equilibrium can be associated with significant health consequences (Hou et al. 2022; Manichanh et al. 2012). Statistically, identifying these specific signals of dysbiosis within raw sequencing outputs is highly demanding due to the inherent high-dimensionality of microbiota compisition data, which typically feature a large number of variables (taxa)

relative to a limited number of available samples (Li 2015). Consequently, understanding the microbiota is no longer solely a biological endeavor, but fundamentally an information extraction challenge.

1.2 Microbiota Characterization: from Ecosystem to Count Matrix

Characterizing microbial communities at the host population level relies on the analysis of biological specimens—such as feces, vaginal swabs, or skin samples—collected from multiple hosts. These samples potentially represent distinct disease cohorts or study groups. This characterization is predominantly achieved via high-throughput DNA sequencing, specifically through targeted amplification of the 16S rRNA gene or shotgun metagenomics (Knight et al. 2018). Following sequencing, the generated DNA fragments (reads) are processed to construct a taxonomic abundance table. In the context of targeted amplicon sequencing, these reads are typically clustered into Operational Taxon Units (OTUs) or Amplicon Sequence Variants (ASVs), whereas shotgun metagenomics employs distinct taxonomic profiling methods. Regardless of the methodology, the resulting data is structured as an $N \times p$ count matrix, where the N rows represent the individual samples and the p columns correspond to the identified taxonomic features. Each cell within this matrix contains the absolute read count assigned to a specific taxon for a given sample. While this numerical abstraction is advantageous, transforming complex ecological ecosystems into tractable mathematical objects, these count matrices inherently exhibit three fundamental statistical properties that must be rigorously addressed in downstream analyses (Knight et al. 2018).

The first prominent property is the high sparsity of microbiota count matrices, characterized by a substantial proportion of zeros. This sparsity arises from two distinct sources. Biologically, it can reflect the true absence of a specific taxon within an individual's microbiota. Technically, it occurs when a bacterium present in the ecosystem is either not captured during sampling or falls below the sequencer's detection threshold due to insufficient read depth (Kaul et al. 2017).

The second fundamental property is that these count matrices typically exhibit overdispersion. Biologically, bacterial abundances are heterogeneous: highly dominant species may be represented by millions of copies, whereas rarer taxa might be detected at very low frequencies, even within a stable ecosystem. Technically, the Polymerase Chain Reaction (PCR) amplification and sequencing processes introduce multiplicative background noise (McLaren et al. 2019). Consequently, the variance of the observed counts consistently exceeds their mean (Sankaran and S. Holmes 2019). This overdispersion has critical statistical implications, notably violating the homoskedasticity assumption required by standard parametric methods, such as ANOVA or Student's t-tests. Rigorous data analysis thus necessitates approaches capable of explicitly modeling this variance or applying appropriate data transformations. Depending on the matrix structure, this involves utilizing probability distributions specifically adapted to extreme variances: the negative binomial distribution for raw counts, zero-inflated models to address excess sparsity, or Beta and Dirichlet distributions to capture the heterogeneity of compositional data (McMurdie and S. Holmes 2014; Paulson et al. 2013; Ferrari and Cribari-Neto 2004; Forbes et al. 2011).

Finally, most microbiota studies rely on analysis protocols that do not control for the multiplicative uncertainty introduced at different sample collection and processing steps (sampling, dilution, DNA extraction, amplification, and sequencing). Consequently, the observed counts are, on average, proportional to the true number of bacteria for each taxon in a sample, but the sample-specific scaling factors linking these observed counts to the true concentrations remain unknown. This has led to the description of microbiotas in terms of *relative abundances* rather than absolute concentrations, and the classification of data generated with such protocols as *compositional*. Count data generated under these conditions can only be reliably interpreted as representing proportions located within a constrained mathematical space, the simplex, rather than true bacterial concentrations (Gloor et al. 2017). It is, however, theoretically and technically possible to estimate absolute abundances through rigorous quantitative tracking at each experimental stage of the protocol. This is notably achieved via the addition of internal standards (*spike-ins*) or by quantifying total bacterial biomass using qPCR. It is expected that future studies requiring absolute quantification will increasingly adopt these alternative protocols (Vandeputte et al. 2017; Tourlousse et al. 2017).

Beyond these mathematical properties, downstream analyses must account for the dynamic nature of the microbiota, whose compositional profile is shaped by numerous intertwined factors (Debelius et al. 2016). Host-related physiological variables (such as age or immune status) alongside environmental drivers (including diet, lifestyle factors, pollutant exposure, and antibiotic use) fundamentally modulate taxonomic colonization and abundance (Lozupone et al. 2012). These biological determinants are further compounded by technical variability, which is introduced by varying sampling protocols, primer selection during DNA amplification, or the specificities of sequencing platforms (Knight et al. 2018).

Collectively, these biological and technical variables constitute the metadata of a study. These metadata variables frequently account for the systematic variation observed within microbiota datasets. Given their influence on the data structure (Debelius et al. 2016), these metadata variables can serve as criteria to stratify samples and divide the study population into defined cohorts of interest (e.g., grouping by disease status, geographic origin, or reproductive condition). For instance, macro-environmental variables, such as geographic origin or shared dietary habits, are often sufficient to induce distinct microbial signatures (Yatsunenko et al. 2012; Zhernakova et al. 2016). Similarly, purely methodological variables, such as the use of varying DNA extraction kits across distinct study centers, can generate technical biases (batch effects) capable of masking true biological signals (Costea et al. 2017; Wesolowska-Andersen et al. 2014).

Ultimately, translating these high dimension, metadata-rich count matrices into robust ecological insights necessitates the application of advanced computational methodologies.

1.3 Capturing Latent Biological Structures using Dimension Reduction Methods.

To facilitate the interpretation of complex microbiota datasets and their relationship with metadata variables, analytical approaches initially sought to reduce the high-dimensionality of count matrices. A standard strategy involved evaluating the global compositional similarity between samples to partition the dataset into lower-dimensional clusters. This

method, known as hard-clustering, assigns each microbiota profile to a single, mutually exclusive category based on its resemblance to a reference signature or dominant taxon. This rigid classification framework led to landmark concepts such as enterotypes for the gut microbiota, which categorize communities into three primary groups (Arumugam et al. 2011; Costea et al. 2018), and Community Sub Typess (CSTs) for the vaginal microbiota. The latter initially distributed samples into five strict groups (Ravel et al. 2011) before being expanded to 13 sub-communities (France et al. 2020).

While these discrete clusters historically aided in correlating broad biological states with metadata, their mathematical rigidity is increasingly recognized as a limitation (Sankaran and S. Holmes 2019). Forcing a complex ecosystem into a single discrete state often fails to capture the continuous and transitional nature of microbial communities. To mitigate this lack of flexibility, an alternative approach consists of examining the data through the lens of mixed-membership partitioning (soft-clustering) (Airoldi et al. 2014). Among these methods, Topic Modeling, initially developed for Natural Language Processing (NLP), appears to offer a particularly relevant conceptual framework (Sankaran and S. Holmes 2019).

In classical textual analysis, Topic Modeling algorithms extract latent semantic structures, referred to as **topics**, by grouping words that frequently co-occur within a corpus: the collection of studied documents. Consequently, a document is no longer classified into a single discrete category. Instead, it is modeled as a mixture of several topics present in varying proportions (Blei et al. 2003).

The transposition of these NLP concepts to the analysis of microbial communities operates through a direct analogy. A biological sample is conceptualized as a document, and the individual taxa (or ASVs) represent the words. Consequently, a latent bacterial community, is defined as a group of taxa that frequently co-occur across multiple samples, corresponds to a topic (Sankaran and S. Holmes 2019). This probabilistic framework allows a single sample to simultaneously contain several distinct sub-communities. Furthermore, it permits a single taxon to belong to different topics in variable proportions. This flexibility more accurately reflects the continuous dynamics of biological ecosystems, where a specific bacterium may occupy distinct ecological niches or participate in varying metabolic networks depending on its local environment.

Several unsupervised learning algorithms facilitate the application of this concept, with the most prevalent being Non-negative Matrix Factorization (NMF) (Wang and Zhang 2013) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003). Fundamentally, these algorithms are dimensionality reduction methods. They project the original $N \times p$ sequencing matrix into a lower-dimensional latent space defined by a limited number of K latent variables. Through this process, the data is structured into two complementary matrices: an $N \times K$ matrix capturing the prevalence of these latent structures for each sample, and a $K \times p$ matrix defining the taxonomic composition of each latent variable.

The specific selection of the LDA model is justified by its capacity to directly accommodate the statistical properties inherent to sequencing matrices. Unlike NMF, which relies on linear algebraic factorization, LDA is a generative probabilistic model whose inference is parameterized by the Dirichlet distribution (Blei et al. 2003; Egger and Yu 2022; Sankaran and S. Holmes 2019). This mathematical distribution is explicitly designed to model variables constrained within a simplex (Forbes et al. 2010b), making it naturally suited to

the compositional nature of microbiota data. Furthermore, its probabilistic framework explicitly accounts for the sparsity and overdispersion of counts. Consequently, LDA represents a reliable analytical tool for inferring true biological signals from complex sequencing datasets (Blei et al. 2003; Sankaran and S. Holmes 2019).

1.4 Testing for Global Cohort Effects via PERMANOVA

While the LDA framework projects the high-dimensional count matrix into a lower-dimensional latent space, this inference process remains strictly unsupervised. The algorithm reconstructs probabilistic topic distributions across samples without incorporating any group-level metadata (Blei et al. 2003). To fulfill the primary objective of this work, which is to assess if *a priori* defined cohorts exhibit systematic differences in their latent topic proportions, a supervised statistical evaluation must be applied *a posteriori*.

However, traditional multivariate procedures like Multivariate Analysis Of Variance (MANOVA) are fundamentally inappropriate for this task. They inherently operate within Euclidean space, a geometry that falsely interprets joint absences (due to sparsity) of taxa as biological similarity. Furthermore, these classical tests are algebraically restricted by high dimensionality ($N \ll p$) and fail to adequately accommodate the strictly compositional nature of microbiota-derived matrices. Consequently, the Permutational Multivariate Analysis of Variance (PERMANOVA) framework is introduced to evaluate whether the grouping covariate accounts for a significant proportion of the variance observed within the inferred latent space (Anderson 2001). By calculating a pseudo-F statistic directly from a chosen ecological distance metric (e.g., Bray-Curtis) and employing distribution-free permutation algorithms, this approach assesses statistical significance while naturally accommodating the sparse and complex properties of microbial ecosystems (Anderson 2017).

Together, the sequential integration of LDA and PERMANOVA constitutes an analytical pipeline for microbiota composition data (Huo et al. 2025).

1.5 Study Overview and Manuscript Structure

Building upon these methodological foundations, the primary objective of this project is to evaluate whether an unsupervised dimensionality reduction approach, combined with a supervised *a posteriori* statistical evaluation, can reliably identify global differences between *a priori* defined cohorts. To achieve this, the analytical pipeline must be assessed against a known ground truth. Thus, microbiota composition data are generated through *in silico* simulations. This synthetic generation guarantees complete knowledge of the underlying ground truth and enables the precise isolation of specific variance parameters. During this generative phase, sample-specific factors, specifically, the cohort assignments, are mathematically linked to the generative process to simulate dependencies with topic prevalence: each cohort is characterized by a distinct topic prevalence profile. This procedure explicitly conditions the generation of the data on the group metadata, thereby enforcing a controlled structural difference among the simulated cohorts. Subsequently, the LDA model is applied to infer the latent thematic structures from the generated count matrices without access to the group labels. Finally, the PERMANOVA framework

evaluates whether these inferred latent proportions exhibit systematic variations that are significantly explained by the original cohort affiliations.

To fulfill these objectives, this work initially details the theoretical and mathematical foundations justifying the use of these probabilistic models in microbial ecology. It subsequently outlines the architecture of the simulation and inference pipeline, before evaluating its reliability and discussing the implications of these results for microbiota cohort studies.

2. State of the Art

2.1 Statistical Properties of Microbiota Composition Data

Datasets generated by high-throughput sequencing of microbial communities exhibit statistical properties that frequently limit the applicability of standard analytical methods. Microbiota abundance matrices are inherently characterized by three major structural constraints: compositionality, sparsity, and overdispersion (Gloor et al. 2017; Sankaran and S. Holmes 2019). Inadequate accommodation of these properties is identified as a primary source of statistical bias and false positive discoveries in microbiota research (Gloor et al. 2017).

2.1.1 Compositionality in Sequencing-Derived Microbiota Composition Data

While the true population dynamics within a natural microbial ecosystem are not inherently compositional, the high-throughput sequencing process fundamentally alters this data structure. Sequencing platforms impose a maximum read capacity, establishing an arbitrary total sequencing depth per sample (Gloor et al. 2017). Consequently, absolute biological abundances are lost; an increase in the read count of one taxon mathematically constrains the available reads for all others. The resulting matrices are thus projected onto a simplex space where only relative abundances hold biological meaning, a mathematical framework formally established by Aitchison (1983).

This structural constraint generates a spurious negative correlation bias among taxa. In the literature, it is widely recognized that this bias invalidates the direct use of standard ecological metrics, such as the Bray-Curtis dissimilarity, on raw count data (Gloor et al. 2017). To circumvent these compositional effects, several normalization strategies are frequently discussed. However, as demonstrated by McMurdie and S. Holmes (2014), classical approaches exhibit severe limitations when applied directly to microbiota data. Rarefaction (random subsampling) artificially discards valid observations, leading to an unnecessary loss of variance and statistical precision (McMurdie and S. Holmes 2014; Gloor et al. 2017). Furthermore, scaling methods initially developed for transcriptomics, such as Trimmed Mean of M-values (TMM) or Differential Expression analysis for Sequencing (DESeq), inherently assume dense matrices (Paulson et al. 2013). Consequently, they remain largely unsuited for microbiota composition profiles, whose mathematical resolution is hindered by a second major structural constraint: data sparsity (Hawinkel et al. 2019).

2.1.2 Sparsity in Microbiota Data: Biological Zeros, Technical Dropouts

A fundamental characteristic of microbiota composition data is extreme sparsity. As detailed by Paulson et al. (2013), while transcriptomic matrices are generally dense, microbiota abundance tables are predominantly populated by zeros. This sparsity originates from two distinct phenomena: **biological** zeros, where a taxon is genuinely absent from the sampled ecosystem, and **technical** zeros (or dropouts), where a taxon is present but remains unrecorded due to the arbitrary sequencing depth bottleneck (Paulson et al. 2013; Gloor et al. 2017).

Consequently, this structural sparsity renders continuous data transformations inadequate. Historically, logarithmic functions were frequently employed, either as simple transformations to moderate variance across taxa (Paulson et al. 2013) or as log-ratio transformations to project compositional data into Euclidean space (Gloor et al. 2017). However, because the $\log(0)$ function is mathematically undefined, these approaches necessitated the addition of arbitrary pseudocounts, typically $\log(x + 1)$ (Kaul et al. 2017). In highly sparse matrices, this practice distorts the underlying data structure: a pseudocount disproportionately impacts shallowly sequenced samples compared to deeply sequenced ones, thereby inducing artificial variation.

As demonstrated by extensive benchmarking studies, applying standard univariate models to sparse, pseudocount-adjusted data yields excessively high Type I error rates. Such analytical strategies frequently misclassify technical dropouts as statistically significant biological absences (Thorsen et al. 2016; Hawinkel et al. 2019).

2.1.3 Overdispersion in Microbiota Count Data: Variance-to-Mean Excess

Beyond sparsity and compositionality, microbiota abundance matrices are fundamentally characterized by extreme structural heterogeneity, or overdispersion. From a statistical perspective, the theoretical baseline for modeling count data is the Poisson distribution, which assumes that the variance equals the mean ($\mathrm{Var}(X) = \mathbb{E}[X]$). However, in microbiota composition data, the observed sample variance of taxon counts consistently and substantially exceeds their empirical mean ($S_X^2 \gg \bar{X}$) (Thorsen et al. 2016; Sankaran and S. Holmes 2019).

This phenomenon, termed overdispersion, systematically violates the assumptions of standard Poisson models and classical linear frameworks. Consequently, applying conventional statistical tests to such data yields diminished power for detecting genuine abundance differences. Furthermore, unadapted modeling strategies frequently misclassify background variance as statistically significant, thereby generating false positive biomarkers (Hawinkel et al. 2019).

To mitigate these analytical artifacts, the robust estimation and simulation of microbiota profiles require probability distributions capable of accommodating this excess variance. In this context, compound frameworks such as the Negative Binomial (NB) or the Dirichlet-Multinomial distribution are indicated to absorb microbiota overdispersion (Blei et al. 2003; I. Holmes et al. 2012; Sankaran and S. Holmes 2019; Symul et al. 2023; Valle et al. 2021).

2.1.4 Methodological Implications for Multivariate Analysis

The combined structural constraints of compositionality, sparsity, and overdispersion highlight a critical methodological bottleneck in microbiota analysis. As previously discussed, relying on classical, unadapted univariate statistics to evaluate taxa individually across samples yields high Type I error rates and unreliable biological interpretations (Hawinkel et al. 2019; Thorsen et al. 2016). These limitations indicate that isolating highly interdependent variables without appropriate compositional transformations remains an inherently limited approach for microbiota composition data (Gloor et al. 2017).

Consequently, analyzing microbiota profiles frequently necessitates a shift from unadapted taxon-by-taxon comparisons toward global, multivariate evaluations. To capture the compositional shifts between different biological conditions, the analytical framework must assess the community structure simultaneously. This requirement has established distance-based multivariate hypothesis testing as a standard analytical baseline in microbial ecology (Robinson et al. 2016; Gloor et al. 2017), enabling the statistical evaluation of global cohort-level differences.

2.2 Differential Analysis and Distance-Based Multivariate Testing

2.2.1 Multivariate Hypothesis Testing Frameworks

As outlined, the inherent constraints of microbiota composition data require analytical frameworks capable of evaluating entire microbial structures simultaneously. To address this need, several multivariate statistical testing methods are available in ecological and bioinformatic research. Classical approaches, such as MANOVA, operate under the assumption of multivariate normality and require dense, continuous data structures, and strictly demand that the number of samples exceeds the number of variables ($N > p$) (Anderson 2001; Anderson and Walsh 2013). Consequently, these parametric frameworks are fundamentally inappropriate for the high-dimensionnal, sparse and overdispersed raw (untransformed) matrices characteristic of microbiota composition (Hawinkel et al. 2019; Sankaran and S. Holmes 2019; Paulson et al. 2013). Alternative non-parametric methods, including Analysis Of Similarities (ANOSIM), rely solely on the ranks of dissimilarities rather than their actual magnitudes, which can result in a loss of statistical precision and an inability to accommodate complex multi-factorial experimental designs (Anderson 2001; Anderson and Walsh 2013).

To circumvent these limitations, PERMANOVA, fundamentally developed by Anderson (2001), has been established as a widely adopted non-parametric framework for community-level differential analysis in microbial ecology (Robinson et al. 2016; Gloor et al. 2017). Unlike standard parametric models, the PERMANOVA framework is flexible and partitions the variance directly within a user-defined dissimilarity space, such as the Bray-Curtis dissimilarity or the Jensen-Shannon Divergence (JSD). This approach allows for the evaluation of global multivariate variations without forcing data transformations to meet unrealistic assumptions of multivariate normality. Instead, any applied data transformations (e.g., square-root or logarithmic) serve as deliberate ecological choices to appropriately weight the contributions of rare versus dominant taxa prior to calculating

the ecological distances (Anderson 2001).

2.2.2 Mathematical Foundations of PERMANOVA test

The statistical mechanics of the semi-parametric PERMANOVA framework rely on the calculation of a pseudo-F statistic. This metric quantifies the ratio of mean among-group variation to mean within-group variation directly from a chosen dissimilarity matrix. Under the null hypothesis (H_0), which postulates the equivalence of group centroids within the space of the chosen dissimilarity measure, the exact probability distribution of this statistic cannot be assumed to follow a standard Fisher-Snedecor F-distribution. This limitation arises from the non-normal properties of ecological data, the inherent interdependence of response variables, and the frequent use of non-Euclidean distances. Consequently, significance is assessed through distribution-free permutation procedures. Assuming the exchangeability of sampling units under a true null hypothesis, sample labels corresponding to categorical covariates (such as cohort affiliations) are randomly shuffled across observations to empirically construct the null distribution of the pseudo-F statistic (Anderson 2001).

While this permutation procedure evaluates true shifts in multivariate centroid positions, its theoretical validity implies multivariate homogeneity of dispersions across groups. Fortunately, empirical simulations demonstrate that PERMANOVA remains robust to heterogeneous dispersions provided the experimental design is balanced (Anderson and Walsh 2013). However, in unbalanced designs, severe differences in group dispersions can confound the analysis. In such specific cases, a statistically significant result may erroneously reflect unequal within-group variances rather than a genuine biological shift in cohort centroids (Anderson and Walsh 2013).

While the PERMANOVA framework mathematically accommodates the extreme sparsity and high dimensionality of microbial counts through appropriate distance metrics (Anderson 2001), evaluating community structure strictly at the individual taxon level often obscures higher-order ecological dynamics and compositional interactions (Gloor et al. 2017). To circumvent the noise inherent to raw feature-level data and to accurately capture the overlapping sub-communities driving cohort differences, these complex profiles can be projected into a reduced latent space. Consequently, this ecological requirement has driven the adoption of probabilistic dimensionality reduction algorithms to complement global distance-based evaluations (Sankaran and S. Holmes 2019; Symul et al. 2023).

2.3 From Clustering to Mixed-Membership Models for Microbiota Dimensionality Reduction

2.3.1 The Limitations of Hard-Clustering Approaches

As established, evaluating microbiota composition strictly at the individual feature level obscures higher-order compositional interdependencies (Gloor et al. 2017). To circumvent this limitation and extract interpretable patterns from high-dimensional abundance matrices, early analytical strategies in microbial ecology relied heavily on strict data simplification (Robinson et al. 2016). As highlighted by France et al. (2020), rather than evaluating the individual abundances of hundreds of taxa simultaneously, assigning a single

categorical label to an overarching microbial profile provides a more tractable framework for cohort-level comparisons and statistical modeling. Consequently, this methodological drive for simplification translated into the widespread application of strict partitioning, or hard-clustering algorithms, to categorize microbial ecosystems into discrete CSTs (France et al. 2020; Ravel et al. 2011) or enterotypes (Koren et al. 2013; Costea et al. 2018).

Methodologically, these hard-clustering frameworks rely on the calculation of pairwise distance or dissimilarity matrices (such as the JSD or UniFrac) between samples, followed by unsupervised clustering algorithms like K-means, Partitioning Around Medoids (PAM) (Koren et al. 2013), or hierarchical clustering (France et al. 2020). The primary objective is to assign each observation to a single, mutually exclusive category, typically representing a community state dominated by specific taxa (Arumugam et al. 2011).

Despite their historical utility, strict partitioning methods present significant ecological and statistical limitations. Fundamentally, hard-clustering imposes artificial boundaries on microbial ecosystems that are inherently distributed along continuous gradients (Koren et al. 2013; Costea et al. 2018). Summarizing a set of observations solely by cluster centroids obscures intra-group variance and the subtle compositional architecture of individual profiles (I. Holmes et al. 2012). Furthermore, these rigid frameworks lack mathematical flexibility: a sample biologically situated at the ecological boundary of two community states is forced into a single category (France et al. 2020; Symul et al. 2023). Consequently, minor technical variations during sequencing or data processing can induce artefactual reclassifications, abruptly shifting a continuous biological profile from one discrete analytical group to another (Koren et al. 2013; France et al. 2020).

2.3.2 The Probabilistic Transition: Dirichlet Multinomial Mixtures (DMM)

To mitigate the core limitations of rigid boundary imposition inherent to classical clustering, analytical strategies in microbial ecology transitioned toward probabilistic frameworks. Originally developed in natural language processing to model heterogeneous word counts across documents, the Dirichlet Multinomial Mixture (DMM) model was formally adapted for microbiota composition data by I. Holmes et al. (2012). Instead of relying on geometric distances, this generative approach explicitly models the underlying biological and technical data-generating mechanisms.

Within this framework, the multinomial distribution mathematically represents the sequencing process itself. It translates true bacterial proportions into discrete reads by modeling stochastic sampling, thereby natively accommodating variations in sequencing depth and extreme data sparsity. Concurrently, the Dirichlet distribution models the natural biological variation among samples belonging to the same metacommunity. Rather than relying on a fixed and rigid centroid, the Dirichlet distribution defines the typical compositional profile of a cluster while controlling for the expected variance and overdispersion around this mean (I. Holmes et al. 2012).

Despite improving the statistical estimation of overdispersed counts, the standard DMM framework operates fundamentally as a hard-clustering model. It assumes the existence of multiple discrete Dirichlet metacommunities but enforces the constraint that each sample originates entirely from one, and only one, of these underlying distributions. Consequently, this mutual exclusivity remains overly restrictive for accurately capturing the continuous,

overlapping reality of biological ecosystems (Pappalardo et al. 2024; Sankaran and S. Holmes 2019).

2.3.3 Topic Modeling: The Mixed-Membership Approach

Because microbial abundances are frequently distributed along continuous gradients rather than forming discrete natural clusters (Koren et al. 2013; Costea et al. 2018), strict partitioning frameworks limit the accurate representation of transition states and overlapping community dynamics (Symul et al. 2023). As highlighted by France et al. (2020) and Symul et al. (2023), many microbial profiles exist in intermediate compositional states between defined CSTs. To address this methodological gap, mixed-membership latent variable models, originally developed for natural language processing, have been adapted to evaluate microbiota composition (Sankaran and S. Holmes 2019). Among these approaches, Topic Models, most notably Latent Dirichlet Allocation (Blei et al. 2003), provide a mathematical architecture capable of resolving these continuous biological structures without imposing artificial boundaries.

Unlike strict partitioning approaches, these models introduce the concept of **mixed-membership**. Within this paradigm, latent sub-communities (or topics) are defined by specific probability distributions over the bacterial taxa. Consequently, samples are no longer assigned to mutually exclusive clusters. Instead, they are evaluated as continuous mixtures of these underlying sub-communities (Blei et al. 2003).

For instance, rather than assigning an observation exclusively to a single category (such as "Cluster 1" or "Cluster 2"), the LDA framework evaluates its profile as a fractional mixture, such as 70% of "Topic 1" and 30% of "Topic 2" (Blei et al. 2003). This latent decomposition does not directly equate to raw sequence counts; rather, these mixture weights are probabilistically inferred from the overall dataset. If a sample contains taxa X, Y, and Z, the corresponding 70/30 ratio is inferred because Topic 1 is mathematically defined by a high probability of generating taxa X and Y (learned from their frequent co-occurrence) whereas Topic 2 is characterized by a higher probability of generating taxon Z (Sankaran and S. Holmes 2019; Pappalardo et al. 2024).

This analytical granularity enables a more precise representation of the observed biological continuum (Symul et al. 2023). In empirical evaluations of vaginal microbiota, modeling samples as continuous mixtures of topics, rather than strictly assigning them to discrete CSTs, provided a more accurate estimation of the overarching community architecture while simultaneously reducing the number of required latent dimensions.

However, transitioning from rigid partitioning to mixed-membership paradigms primarily addresses the biological requirement for continuous representation. To fully capture microbiota dynamics, the chosen dimensionality reduction framework must also accommodate the statistical constraints (sparsity, overdispersion and compositionality) inherent to high-throughput sequencing data. This dual requirement necessitates a critical evaluation of the available mixed-membership algorithms.

2.3.4 Evaluating Soft-Clustering Models: Algebraic (NMF) vs. Probabilistic Frameworks (LDA)

Non-negative Matrix Factorization (NMF) is one of the most used mixed-membership methods for dimensionality reduction (Wang and Zhang 2013). Its principle relies on the factorization of a large abundance matrix V into two lower-dimensional matrices: a topic matrix (W) and a proportion matrix (H). However, in its classical form, NMF remains a purely algebraic and geometric framework. It implicitly assumes that errors are distributed normally, as the minimization of geometric distances inherently models additive Gaussian noise (Wang and Zhang 2013). Consequently, this architecture does not natively accommodate the biological nature of sequencing counts: zeros are treated as absolute absences (Wang and Zhang 2013), and extreme variance is mathematically captured as relevant structural information rather than artefactual overdispersion generated by the sequencing process. Furthermore, as recently highlighted by Duncan et al. (2025), unmodified NMF fails to address the statistical issues inherent to compositional data, while standard compositional workarounds (such as log-ratio transformations) directly violate the strict non-negativity constraint required for its mathematical convergence.

To mitigate these limitations, matrix factorization evolved toward probabilistic generative models, such as the Gamma-Poisson (GaP) framework (Canny 2004). By employing discrete probability distributions (such as the Poisson or Negative Binomial distribution) rather than geometric distances, these frameworks integrate the stochastic nature of sequence counts. They provide a mathematical architecture capable of accommodating the overdispersion and extreme sparsity that frequently characterize microbiota composition data (Sankaran and S. Holmes 2019).

While probabilistic models based on factorization (GaP) and those based on mixed-membership (such as LDA) appear conceptually adapted to microbiota constraints, their empirical performances diverge substantially. In a comprehensive benchmarking study on microbiota composition data, Sankaran and S. Holmes (2019) indicated a consistent statistical advantage of the LDA model over the GaP model:

- **Mathematical identifiability:** The GaP framework exhibits a structural lack of mathematical identifiability. Because its prior distributions are not sufficiently constrained, the inference process frequently fails to define the appropriate scale of factorization. Conversely, the LDA architecture is strictly constrained by its Dirichlet priors, effectively bypassing this limitation.
- **Estimation accuracy:** Empirical simulations demonstrate that parameter estimation errors for LDA are systematically lower than those for GaP. Furthermore, LDA accuracy consistently scales with increased sequencing depth and larger sample sizes.
- **Algorithmic convergence:** While inference procedures for LDA converge with high stability, GaP implementations frequently fail to converge or tend to stabilize around artefactual local optima.

Given this recurrent mathematical instability, Sankaran and S. Holmes (2019) ultimately excluded the GaP model from their empirical evaluations on real cohort data, effectively establishing LDA as a more statistically accurate framework for modeling microbiota dynamics.

2.4 Theoretical Model: Latent Dirichlet Allocation (LDA)

LDA, introduced by Blei et al. (2003), is a hierarchical generative probabilistic model originally designed to infer latent structures within discrete data collections, primarily text corpora. Initially developed for NLP, the model evaluates each document as a fractional mixture of topics, where each topic is strictly defined as a probability distribution over a vocabulary. This mathematical formalism is directly transposed to microbial ecology: each biological sample represents a collection of sequencing reads, and these discrete reads are stochastically generated by underlying functional or ecological bacterial sub-communities, which correspond to the topics (Sankaran and S. Holmes 2019).

2.4.1 From text corpora to sequencing reads: adapting LDA nomenclature to microbiota data

To formalize the generative model, the standard LDA nomenclature (Blei et al. 2003) is adapted to accommodate microbial abundance data (Sankaran and S. Holmes 2019):

- **Word** (w): The fundamental discrete unit of data. Mathematically, a word is an element belonging to a vocabulary of size V, represented by a V-dimensional unit vector. For the v-th word of the vocabulary, the component is $w^v = 1$, while $w^u = 0$ for $u \neq v$.
- **Document** ($\mathbf{w}$): A sequence of N words, denoted as $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n represents the n-th word of the sequence.
- **Corpus** (Dataset): A collection of M documents, denoted as $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

Within the analytical framework of this study, the word corresponds to the assigned taxon (or sequencing read), the document corresponds to the individual biological sample, and the corpus corresponds to the entire experimental dataset.

2.4.2 The Dirichlet Distribution: Simplex Support and Proportional Data Generation

A fundamental characteristic of the LDA framework relies on modeling the fractional proportions of latent sub-communities (topics) within a sample (Blei et al. 2003; Sankaran and S. Holmes 2019). The topic prevalence vector, denoted as θ, is modeled by a Dirichlet distribution parameterized by a K-dimensional hyperparameter vector α:

$$\theta \sim \text{Dir}(\alpha) \tag{2.1}$$

The Dirichlet distribution is a continuous multivariate probability distribution whose support is a $(K-1)$-simplex. Consequently, a single realization from this distribution yields a prevalence vector whose components are non-negative and sum to one ($\sum_{k=1}^{K} \theta_k = 1$).

This structural mathematical constraint is optimally suited for generating proportional data. The probability density function on the simplex is defined by:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} \tag{2.2}$$

2.4.3 The LDA Generative Process and Joint Distribution

The LDA framework outlines a stochastic generative process for each document $\mathbf{w}$ within a corpus D (Figure 2.1). The steps are defined as follows (Blei et al. 2003; Sankaran and S. Holmes 2019):

1. Sample the document depth: $N \sim \text{Poisson}(\xi)$.
2. Sample the topic prevalence vector for the document: $\theta \sim \text{Dir}(\alpha)$ (Eq. 2.1).
3. For each of the N words w_n:
   (a) Sample a topic assignment: $z_n \sim \text{Multinomial}(\theta)$.
   (b) Sample a word w_n from $p(w_n|z_n, \beta)$, representing a multinomial probability conditioned on the assigned topic z_n.

This standard architecture relies on two simplifying assumptions. First, the dimensionality K of the Dirichlet distribution and the topic variable z is considered known and fixed. Second, the word probabilities are parameterized by a $K \times V$ matrix β, where $\beta_{kj} = p(w_j = 1|z_k = 1)$. For this base formulation, β is treated as a fixed parameter to be estimated; this constraint is subsequently relaxed in Section 2.4.4.

Given the parameters α and β, the joint distribution of a topic prevalence vector θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is formalized as (Blei et al. 2003):

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2.3}$$

where $p(z_n|\theta)$ is equal to θ_k for the unique index k such that $z_{nk} = 1$.

By integrating Equation 2.3 over θ and summing over $\mathbf{z}$, the marginal distribution of a document is obtained:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) \mathrm{d}\theta \tag{2.4}$$

The total probability of the corpus D is then the product of the marginal probabilities of each document (Blei et al. 2003):

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) \mathrm{d}\theta_d \tag{2.5}$$

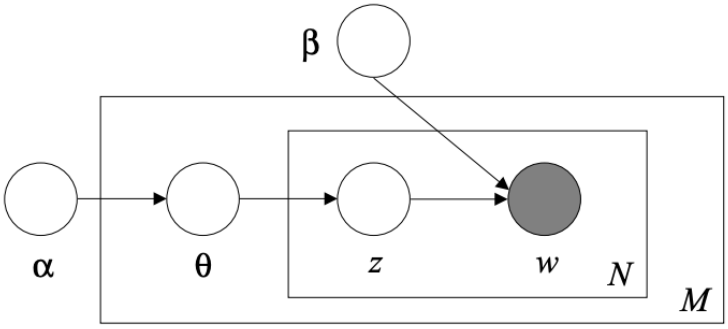


Figure 2.1: **Plate diagram of the standard Latent Dirichlet Allocation (LDA) model.**
The outer plate represents the corpus of M documents (samples), while the inner plate represents the N words (reads) within each document. White nodes denote latent variables and parameters, whereas the shaded node w represents the observed discrete data (observed word / taxon). Specifically, the hyperparameter α parameterizes the Dirichlet prior for the the document-level topic prevalence vector θ. At the word level, z denotes the latent discrete topic assignment responsible for generating the observed word w, conditioned on the fixed topic-word distribution matrix β. Adapted from Blei et al. (2003).

2.4.4 Smoothed LDA: Placing a Dirichlet Prior on Topic Composition (β)

In the standard formulation of LDA, the topic composition matrix β is treated as a fixed parameter. To prevent overfitting (specifically, the assignment of zero probabilities to unobserved taxa during inference) the smoothed model introduces a second Dirichlet prior (Blei et al. 2003).

In this smoothed architecture (Figure 2.2) the topics become random variables. Each topic distribution β_k is drawn from a Dirichlet distribution parameterized by a V-dimensional hyperparameter vector η:

$$\beta_k \sim \text{Dir}(\eta) \tag{2.6}$$

The joint probability applied to the entire corpus D then incorporates this smoothing of the taxon distributions:

$$p(\mathbf{W}, \mathbf{Z}, \Theta, \beta | \alpha, \eta) = \underbrace{\left[\prod_{k=1}^{K} p(\beta_k|\eta)\right]}_{\text{Part 1: Word Smoothing}} \times \underbrace{\prod_{d=1}^{M} \left(p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|\beta_{z_{dn}}) \right)}_{\text{Part 2: Document Structure}} \tag{2.7}$$

Due to its enhanced capacity to accommodate extreme sparsity and high-dimensional biological vocabularies, this smoothed framework represents the standard implementation in contemporary microbiota analyses (Sankaran and S. Holmes 2019). Consequently, this specific architecture constitutes the core analytical model deployed in the present study.

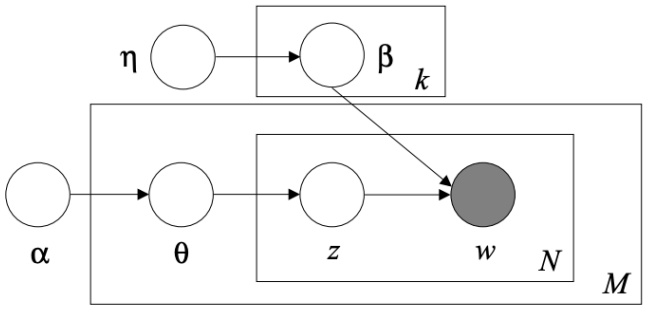


Figure 2.2: **Plate diagram of the smoothed Latent Dirichlet Allocation (LDA) framework.**
The model is represented by three plates: the outer M plate represents the corpus of documents (samples), the inner N plate represents the words (reads) within each document, and a distinct K plate represents the latent topics. White nodes denote latent variables and parameters, whereas the shaded node w represents the observed discrete data (observed word / taxon). Specifically, the hyperparameter α parameterizes the Dirichlet prior for the document-level topic prevalence vector θ, and z denotes the latent discrete topic assignment responsible for generating the observed word w. Furthermore, a second Dirichlet prior η is introduced to parameterize the topic distributions β_k inside the K plate, preventing the assignment of zero probabilities to unobserved taxa during inference. Adapted from Blei et al. (2003).

2.4.5 Posterior Inference: Intractability of the Marginal Likelihood and Approximate Methods

Inference consists of inverting the generative process: starting from the observed variables (the occurrences of taxa $\mathbf{w}$), the objective is to determine the posterior distribution of the latent variables θ (the topic prevalence vector per sample) and $\mathbf{z}$ (the topic assignments for each sequencing read):

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{2.8}$$

The exact calculation of this posterior distribution depends on the denominator $p(\mathbf{w}|\alpha, \beta)$, which represents the marginal probability of the observations (Eq. 2.4). This term is mathematically intractable. The intractability arises from the necessity to compute a continuous integral over the simplex for θ, strongly coupled with a discrete summation over an exponentially large number of possible topic assignments (K^N possible configurations for $\mathbf{z}$) (Blei et al. 2003).

To overcome this analytical bottleneck, exact inference is circumvented through approximate inference frameworks. Two main families of approximation are commonly deployed in the literature:

- **Stochastic approximation (Markov Chain Monte Carlo (MCMC)):** Methods such as Collapsed Gibbs Sampling iteratively sample from conditional distributions to empirically approximate the posterior distribution.

- **Deterministic approximation:** Methods such as Variational Inference (VI) reformulate the inference as an optimization problem, approximating the true posterior with a simpler, parameterized distribution.

Given the substantial computational cost associated with stochastic methods (such as MCMC) on large, high-dimensional microbiota datasets, deterministic approaches are often preferred. Consequently, the Variational Expectation-Maximization (VEM) algorithm is specifically selected for the analytical framework of this study to optimize computational efficiency (Sankaran and S. Holmes 2019).

2.4.6 Variational Expectation-Maximization (VEM) Inference

The VEM method is an iterative procedure that alternates two phases until convergence, maximizing the log-likelihood of the observed data. For a given sample $\mathbf{w}$, consisting of N reads, this log-likelihood is mathematically expressed as (Grün and Hornik 2011):

$$\ell(\alpha, \beta) = \log p(\mathbf{w}|\alpha, \beta) = \log \int \left\{ \sum_{\mathbf{z}} \left[\prod_{n=1}^{N} p(w_n|z_n, \beta) p(z_n|\theta) \right] \right\} p(\theta|\alpha) \mathrm{d}\theta \tag{2.9}$$

However, the coupling between the latent variables θ (prevalence) and β (composition) within the summation renders the exact calculation of this integral intractable within a reasonable timeframe. The VEM method circumvents this problem by maximizing not the exact likelihood, but a tight lower bound of it (Grün and Hornik 2011).

This deterministic framework alternates two iterative phases until convergence:

1. **The E (Expectation) step:** The posterior probabilities of the latent variables (the topic prevalence vectors and topic assignments) are estimated by considering the global model parameters as temporarily fixed.
2. **The M (Maximization) step:** These variational estimates are subsequently utilized to update and optimize the global model hyperparameters (such as α and β).

By iteratively applying these two steps, the VEM framework provides a computationally efficient approximation, enabling the application of complex generative models like LDA to high-dimensional biological matrices (Sankaran and S. Holmes 2019).

2.5 Integrating Covariates into LDA: the Dirichlet-Multinomial Regression (DMR) framework

Validating a mixed-membership inference pipeline requires a generative architecture that explicitly encodes latent community structures (Sankaran and S. Holmes 2019; Symul et al. 2023). Standard simulation frameworks that independently model marginal taxon distributions (e.g., using Zero-Inflated Negative Binomial (ZINB) models) (Hawinkel et al. 2019) do not satisfy this requirement, as they generate counts without underlying topic structure for the algorithm to recover. Therefore, the simulation framework developed here directly extends the LDA generative process.

In the standard LDA framework, the hyperparameter α, which governs the prior distribution of the topic prevalence, is static and shared across the entire corpus (Blei et al. 2003). Consequently, the generative process is strictly unsupervised and does not account for overarching sample metadata (such as group assignments).

Yet in microbiota studies, such metadata typically encode physiologically or experimentally relevant groupings (Debelius et al. 2016) (e.g., phenotypic status, dietary intervention, or other cohort membership) whose association with sub-community structure constitutes the primary biological question of interest. To theoretically link discrete compositional data to external continuous or categorical variables, specific extensions to the generative framework have been developed.

Several probabilistic architectures exist to bridge latent topics and covariates. For instance, the Structural Topics Model (STM) (Roberts et al. 2014) allows metadata to influence both the topic prevalence (α) and the vocabulary distribution within these topics (β). Conversely, models such as Latent Dirichlet Allocation with covariates (LDAcov) (Valle et al. 2021) condition the total absolute sequence counts assigned to each topic based on metadata.

However, in the context of microbial ecology, the taxonomic signature (the β matrix) of a sub-community generally must remain fixed and universal to enable rigorous cross-cohort comparisons (Sankaran and S. Holmes 2019). Furthermore, absolute sequencing abundance is frequently considered a technical artifact rather than a direct biological measurement. To adhere to these structural constraints, the Dirichlet Multinomial Regression (DMR) framework, introduced by Mimno and McCallum (2012), provides a suitable mathematical formulation.

Instead of modifying the overarching topic composition, the DMR framework adopts a conditional approach. It directly conditions the prior distribution of the topic prevalence on the observed characteristics of each document via a log-linear link function. Mathematically, the DMR model associates an observed metadata vector x_d (e.g., categorical cohort indicators) with each document d. For each topic k, the model defines a parameter vector λ_k of the same dimension as x_d. The global concentration parameter α is thereby substituted by a document-specific vector α_d, computed according to the following equation (Mimno and McCallum 2012):

$$\alpha_{d,k} = \exp(x_d^\top \lambda_k) \tag{2.10}$$

Through this exponential formulation, the concentration parameters are mathematically guaranteed to remain non-negative, while the λ_k coefficients can assume negative values to model a suppressive effect of a covariate on a specific sub-community.

Once the components $\alpha_{d,k}$ are computed to construct the sample-specific prior vector α_d, the generative process resumes exactly as in the standard LDA model: the topic prevalence vector (θ) for the sample is sampled from $\theta_d \sim \text{Dir}(\alpha_d)$. The DMR framework thus offers a mathematically sound mechanism to integrate metadata into a generative model by shifting the distributions of topic proportions while maintaining a unified taxonomic vocabulary. By encoding specific group assignments as covariates, this formulation naturally generates the distinct structural variations required to simulate differentiated cohorts.

2.6 Synthesis and Problem Statement

The statistical properties of microbiota sequencing data—namely compositionality, extreme sparsity, and high overdispersion—severely constrain the choice of analytical frameworks. Traditional distance-based methods and hard-clustering algorithms often fail to capture the continuous, overlapping nature of microbial ecosystems. To address these limitations, mixed-membership models, particularly the smoothed LDA framework, provide a mathematically suitable generative architecture. By modeling discrete sequence counts as mixtures of latent sub-communities, LDA explicitly accounts for the structural constraints of high-dimensional microbiota matrices.

Despite these theoretical advantages, LDA remains strictly unsupervised during inference. While generative extensions such as the DMR framework offer a mechanism to simulate distinct cohort profiles by conditioning topic prevalence on metadata, the empirical reliability of the inverse analytical pipeline requires systematic validation. Specifically, the capacity of standard, unsupervised LDA to accurately reconstruct latent spaces from data generated with metadata-driven constraints, such that downstream statistical tests (e.g., PERMANOVA) can reliably detect overarching group variations, requires further evaluation.

The impact of sample size, biological effect amplitude, and residual overdispersion on both inference fidelity and statistical power has yet to be thoroughly characterized. Building on these theoretical foundations, the present study addresses this gap through a controlled, multifactorial simulation framework.

3. Research Objectives

The primary objective of this study is to evaluate the capacity of an analytical pipeline, based on the LDA model, to infer overarching differences between cohorts defined by their topic prevalence profile (inspired from the DMR generative process). To validate this approach, a simulation framework generating synthetic data is developed. In this model, the assignment of samples to the different groups is defined *a priori*. Through a full factorial design, this study jointly evaluates:

1. The fidelity with which the LDA algorithm reconstructs the simulated latent matrices.
2. The statistical power of the PERMANOVA test, applied to the sample-level topic prevalence matrix, to demonstrate that these groups possess significantly distinct centroids, thereby confirming an overarching difference between the cohorts.

To address this primary objective, the methodological approach is structured around three specific objectives:

1. **Develop a simulation framework for microbiota composition sequencing data.**
   This involves designing a pipeline generating a synthetic count matrix that reproduces the characteristics of microbiota sequencing (sparse, overdispersed, and compositional matrix). This model allows the construction of known cohorts by conditioning the prevalence probabilities of the sub-communities (topics) on specific cohort membership: each cohort thus possesses a distinct topic prevalence profile.
2. **Evaluate the reliability of parameter inference by the LDA algorithm.**
   This second objective aims to quantify the capacity of the LDA to faithfully reconstruct the simulated latent matrices (composition (β), prevalence (Γ), and taxon observation probabilities($P = \Gamma \times \beta$)). Through a full factorial design, Generalized Linear Models (GLMs) will evaluate the impact of three simulation parameters (cohort size, biological effect amplitude, and residual variance) on the reconstruction scores.
3. **Measure the reliability of the PERMANOVA test to evaluate the cohort effect.**
   This final objective aims to determine whether the PERMANOVA test, applied to the prevalence matrix inferred by the LDA, can detect the cohort effect. By leveraging the full factorial design, GLM will model test significance (derived from the p-values obtained in the experimental design) to evaluate the impact of the three simulation parameters on statistical power (capacity to detect the effect) and on the control of the false positive rate.

4. Materials and Methods

To address the objectives defined in the previous chapter, this section outlines the technical and methodological framework developed to benchmark the mixed-membership inference pipeline and evaluate its capacity to infer overarching cohort variations. To ensure a rigorous and transparent evaluation, the methodology is structured into four interconnected sections.

First, the *in silico* generative model is detailed, describing the mathematical steps used to construct synthetic, biologically plausible microbiota count matrices embedded with metadata-driven cohort structures.

Second, the experimental design of the simulation study is established, defining the invariant baseline configuration and the full factorial space under which the analytical boundaries are tested.

Third, the comprehensive analytical pipeline is presented. This includes the unsupervised LDA inference, the algorithmic resolution of the stochastic label switching phenomenon, and the subsequent application of the PERMANOVA test to evaluate cohort effects.

Fourth, the evaluation framework is outlined, detailing the quantitative metrics used to assess inference fidelity and statistical power, followed by the formal modeling of these performance outcomes via GLM.

4.1 *In-Silico* Generative Model for Microbiota Community Data

To evaluate the capacity of the unsupervised LDA model to accurately infer latent spaces that preserve the signatures of overarching cohort structures, a comprehensive simulation pipeline is developed. This generative framework is designed to produce *in silico* count data that faithfully reproduce the technical constraints (e.g., extreme sparsity, overdispersion) and biological complexities of microbiota composition data. Furthermore, rather than generating a homogeneous dataset, the pipeline integrates categorical metadata to construct well-defined *a priori* cohorts. By manipulating the latent topic prevalence across these simulated groups, the pipeline provides the ground truth required to benchmark the subsequent inference and statistical testing steps.

4.1.1 Conditioning Topic Prevalence on Cohort Membership in the Simulation Pipeline

In the context of microbial epidemiology, the objective of an observational study is rarely to model a static cohort, but rather to identify structural variations in the microbiota associated with a specific phenotypic state (e.g., disease status, dietary intervention, geography). To evaluate the capacity of the LDA model to preserve the signal of these

biological signatures within its latent space, it is essential to integrate a group structure into the simulation process.

To translate the theory of the DMR model into this simulation environment, the pipeline is adapted to condition the topic prevalence on a categorical metadata variable (e.g., cohort membership). Specifically, the theoretical equation of the DMR is implemented via the construction and matrix multiplication of two specific matrices.

4.1.1.1 The Design Matrix (X)

The purpose of integrating this covariate is to separate the samples into several distinct cohorts, each possessing its own overarching microbial signature. The effect of the covariate solely weights the prevalence of topics between cohorts: each group consists of the exact same latent topics (which are themselves composed of the same taxa) but the topic proportions differ structurally across groups.

The first algorithmic step consists of encoding this group membership as a design matrix via a dummy coding method. Group 1 (G_1) is defined as the baseline reference group. For each sample i, a row vector of size G (where G is the total number of groups) is created:

- The first component (the intercept) always receives a value of 1.
- If the sample belongs to a target group g (where $g > 1$), the corresponding column receives a value of 1, while all other group columns receive a 0.

The vertical aggregation of these row vectors forms the global design matrix X of dimension $N \times G$ (generated via the custom `generate_id_matrix.R` function).

4.1.1.2 Calculation of Conditional Prevalence

To determine the sample-specific concentration parameter α_p (representing the prior prevalence) specific to a topic k for a sample i, the information from the design and weight matrices is combined. To modulate the intensity of the biological effect across the different simulation scenarios, a global Amplitude (A) parameter is added to the log-linear link function:

$$\alpha_{p,i,k} = \exp\left(W_{k,1} + A\sum_{g=2}^{G} W_{k,g}X_{i,g}\right) \tag{4.1}$$

where:

- $\alpha_{p,i,k} \in \mathbb{R}^+$: The non-negative concentration parameter (scalar) dictating the weight of topic k in the Dirichlet prior for sample i. The set of these K parameters forms the vector $\boldsymbol{\alpha}_{p,i}$.
- $W_{k,1} \in \mathbb{R}$ represents the intercept, i.e., the baseline prevalence coefficient of topic k in the reference cohort (G_1).
- $X_{i,g} \in \{0, 1\}$ is the indicator variable designating the sample's membership in the target group g.
- $W_{k,g} \in \mathbb{R}$ is the regression coefficient capturing the specific effect of group g on the prevalence of topic k.

- A is the scaling amplitude parameter used in the experimental design.

In this equation, the biological interpretation of a regression coefficient $W_{k,g}$ is direct, as its effect operates exponentially on the baseline prevalence:

- $W_{k,g} < 0$: Topic k is depleted in the target group (inhibitory effect).
- $W_{k,g} = 0$: The group has no influence on the topic prevalence ($e^0 = 1$).
- $W_{k,g} > 0$: Topic k is enriched in the target group (promoting effect).

Once the set of vectors $\boldsymbol{\alpha}_{p,i}$ is calculated for all samples, it is passed to the core generative model to compute the final count data.

4.1.2 Hierarchical Generative Model

The generative process follows a hierarchical structure to produce four distinct matrices:

1. **The sub-community prevalence matrix (Γ):** It represents the document-topic proportions [1]. It defines the true biological proportion of each latent sub-community within a sample. It is composed of N row vectors γ_i, forming an $N \times K$ matrix. The individual components $\gamma_{i,k}$ represent the specific proportion of topic k, and the sum of each matrix row must strictly equal 1. Furthermore, the proportions for a sample i are drawn from an asymmetric Dirichlet distribution parameterized by the dynamically calculated vector $\boldsymbol{\alpha}_{p,i}$ (derived from Eq. 4.1):

$$\gamma_i \sim \text{Dir}(\boldsymbol{\alpha}_{p,i}) \tag{4.2}$$

2. **The topic composition matrix (B):** It represents the β matrix in the standard theory of the LDA model (Blei et al. 2003; Fukuyama et al. 2021). It defines the distribution of the F taxa within each topic. It is composed of K row vectors β_k, forming a $K \times F$ matrix. To explicitly model a biological complexity gradient among the microbiota sub-communities (a specific methodological addition of this pipeline) K concentration scalars (one for each topic) are first drawn independently from a Beta distribution and sorted in ascending order:

$$\alpha_{c,k} \sim \text{Beta}(1, \eta) \quad \text{for } k \in \{1, \ldots, K\} \text{ (sorted)} \tag{4.3}$$

   Each scalar $\alpha_{c,k}$ is then used to parameterize a symmetric Dirichlet distribution across all F taxa for the corresponding topic k:

$$\beta_k \sim \text{Dir}(\alpha_{c,k} \cdot \mathbf{1}_F) \tag{4.4}$$

   where $\mathbf{1}_F$ is a vector of ones of length F. Because the concentration parameters $\alpha_{c,k}$ are sorted, this explicitly enforces an alpha-diversity gradient within the latent space (Willis 2019). This design leverages the fundamental mathematical property of the Dirichlet prior, where the magnitude of the concentration parameter dictates the sparsity of the drawn distributions (Wallach et al. 2009): topics parameterized with

[1] While this matrix is canonically denoted as the Θ matrix in the foundational LDA literature (Blei et al. 2003), it is denoted here as Γ. This aligns with the mathematical nomenclature for the variational parameters used to approximate the posterior distribution of θ, as utilized in subsequent literature (Fukuyama et al. 2021; Grün and Hornik 2011) and the simulation codebase.

a low $\alpha_{c,k}$ simulate sparse sub-communities strongly dominated by a few specialist taxa, while topics with a higher $\alpha_{c,k}$ simulate more complex, balanced networks of generalist species.

3. **The taxon observation probability matrix (P):** The matrix product of the two previous latent matrices (G and B) . It represents the ideal and exact biological proportion of each taxon in each sample prior to the introduction of sequencing noise. It is composed of N vectors of size F, forming an $N \times F$ matrix. The columns (taxa) are structurally ordered in descending order according to their overall simulated abundance.

$$P = \Gamma \times B \tag{4.5}$$

4. **The count matrix (C):** To simulate the sequencing step, a sequencing depth (library size, D_i) is assigned to each sample according to a log-normal distribution. This is a widely adopted, empirically justified approach used to reproduce the technical variability inherent in the preparation of sequencing libraries, notably the multiplicative biases linked to PCR amplification steps (Ma et al. 2021). The parameters of this distribution are empirically calibrated (log-mean = 9, log-sd = 0.3) to accurately reflect the typical depths observed in real-world biological cohorts.

   Based on these depths (D_i) and the taxon observation probabilities (P), the final read counts can be generated according to three mathematical frameworks:

   - **Multinomial model (MULT):** The foundation of theoretical LDA, modeling simple sampling noise without overdispersion. The count matrix is obtained via a standard Multinomial distribution.
   - **Negative Binomial model (NB):** The primary generative model used in this study to simulate biologically plausible microbiota data. Unlike the Multinomial model, it introduces an extrinsic size parameter (ϕ) that allows the variance to exceed the mean, mimicking the overdispersion characteristic of real sequencing data (Valle et al. 2021). This distributional mismatch (NB-generated counts fed into a Multinomial-assuming LDA) constitutes the ecologically realistic condition under which the inference pipeline is evaluated. The variance structure is governed by: Mathematically, the introduction of this parameter allows the variance to uncouple from and exceed the mean, a property governed by the following equations:

$$E[X] = \mu \tag{4.6}$$

$$Var(X) = \mu + \frac{\mu^2}{\phi} \tag{4.7}$$

     where $\mu = D \times P$
   - **Zero-Inflated Negative Binomial (ZINB):** An extension implemented in the pipeline that allows, in addition to overdispersion, the forcing of an inflation of structural zeros (true biological absences) via an additional probability parameter ω.

4.2 Experimental Design of the Simulation Study

Following the definition of the generative process, an experimental design is established to assess the analytical pipeline against the primary research objectives. Specifically, the design aims to evaluate the inference fidelity of the latent parameters (reconstruction accuracy) and the statistical reliability of the PERMANOVA test applied to the inferred prevalence matrix (Γ). To achieve this, the simulation engine is utilized to generate microbiota count data under varying ecological and structural conditions. This section outlines the architecture of this testing environment, first detailing the invariant baseline configuration, and subsequently defining the variable parameter space explored within a full factorial design.

4.2.1 Invariant Parameters and Baseline Configuration

To strictly isolate the effects of the variables of interest, a foundational set of parameters is kept constant across the different scenarios. These parameters establish a standardized ecological framework:

- **Number of cohorts (G, `n_groups`):** Fixed at **3**.
  The number of levels of the categorical covariate (distinct predefined cohorts).
- **Taxonomic richness (F, `n_taxa`):** Fixed at **100**.
  The total number of taxa present in the simulated ecosystem. All samples share the same taxa, but their relative proportions fluctuate.
- **Number of latent topics (K, `n_topics`):** Fixed at **5**.
  The number of latent sub-communities (topics) constituting the samples. All samples share the same topics, but their overarching proportions fluctuate.
- **Composition prior (η, `prior_composition`):** Fixed at **5**.
  The hyperparameter controlling the sparsity of the composition matrix B. Parameterizing the underlying Beta distribution ($\text{Beta}(1, \eta)$). It strongly skews the drawn concentration parameters ($\alpha_{c,k}$) toward lower values (< 1). As dictated by the mathematical properties of the Dirichlet prior, these low values enforce a high degree of structural sparsity, yielding sub-communities dominated by a restricted subset of specialist taxa.
- **Weight matrix (W):** The structure of the associations between groups and topics is fixed according to a predefined biological signature. This choice ensures the evaluation of the pipeline's capacity to reconstruct an identical complex structure across varying noise conditions (Table 4.1).

Table 4.1: **Predefined Weight Matrix (W) driving the biological group effects.** The intercept ($W_{k,1}$) dictates the baseline log-prevalence of each topic in the reference cohort (Group 1). The subsequent columns represent the regression coefficients applied to target groups. Positive values indicate an enrichment effect on the topic's relative abundance, while negative values indicate a depletion effect.

Latent Topic (k)	**Intercept (Group 1)** ($W_{k,1}$)	**Effect Group 2** ($W_{k,2}$)	**Effect Group 3** ($W_{k,3}$)
Topic 1	1.0	-1.0	-1.0
Topic 2	0.0	1.0	0.0
Topic 3	0.0	0.0	0.5
Topic 4	0.0	0.5	0.5
Topic 5	0.0	0.0	0.5

- **Data generation model:** The exclusive use of the NB emission model is implemented, justified by its fundamental capacity to capture the overdispersion inherent to biological count data.

4.2.2 Variable Parameters and Full Factorial Design

A systematic full factorial design is implemented to rigorously evaluate the pipeline across a vast array of conditions, yielding 5,000 distinct simulation datasets: 100 unique theoretical scenarios, each independently replicated 50 times (Figure 4.1). Three structural levers of variability are systematically actuated:

1. **Group effect amplitude (A):** Five levels.
   Tested over a dynamic range from **0 to 1** $(0, 0.25, 0.5, 0.75, 1)$. The lower bound ($A = 0$) serves as the null hypothesis to evaluate the False Positive Rate (Type I error). Conversely, the value of 1 marks a strong, saturating biological signal, selected to guarantee an expected statistical power nearing 100% under optimal testing conditions.

2. **Cohort size (N_g):** Four levels.
   Tested over a range from **10 to 100** $(10, 25, 50, 100)$ samples per predefined group. These tiers reflect the standard distribution of empirical microbiome studies, ranging from underpowered pilot studies ($N_g = 10$) to large-scale observational cohorts ($N_g = 100$).

3. **Size parameter (ϕ):** Five levels.
   A critical parameter of the NB distribution, tested over a broad range from **0.5 to 50** $(0.5, 1, 5, 10, 50)$. Mathematically acting as an inverse-dispersion parameter in the variance function ($Var = \mu + \mu^2/\phi$)[2], low values (0.5 and 1) generate extreme structural overdispersion typical of noisy biological data. Conversely, high values

[2]In the statistical literature, the coefficient controlling the variance term is alternately defined as a dispersion parameter (α, where $Var = \mu + \alpha\mu^2$, such that higher values increase variance) or a size parameter (ϕ, where $Var = \mu + \mu^2/\phi$, such that higher values decrease variance). The size parameterization is explicitly utilized here to directly align with its native implementation in the `rzinb` package used within the simulation codebase.

(10 and 50) suppress this overdispersion, forcing the distribution to strictly converge towards a Poisson behavior representing an idealized scenario of low variance.

Each combination of these three factors is independently replicated **50 times**. To guarantee the strict reproducibility of the stochastic generative process across these iterations, a random seed management strategy indexed to the specific replicate number (`set.seed(1234 + rep)`) is systematically applied. This replication threshold is calibrated to ensure the asymptotic convergence of the statistical power and Type I error estimates, while maintaining computational feasibility. These simulation parameters are subsequently strictly treated as categorical variables for the downstream statistical modeling in R.

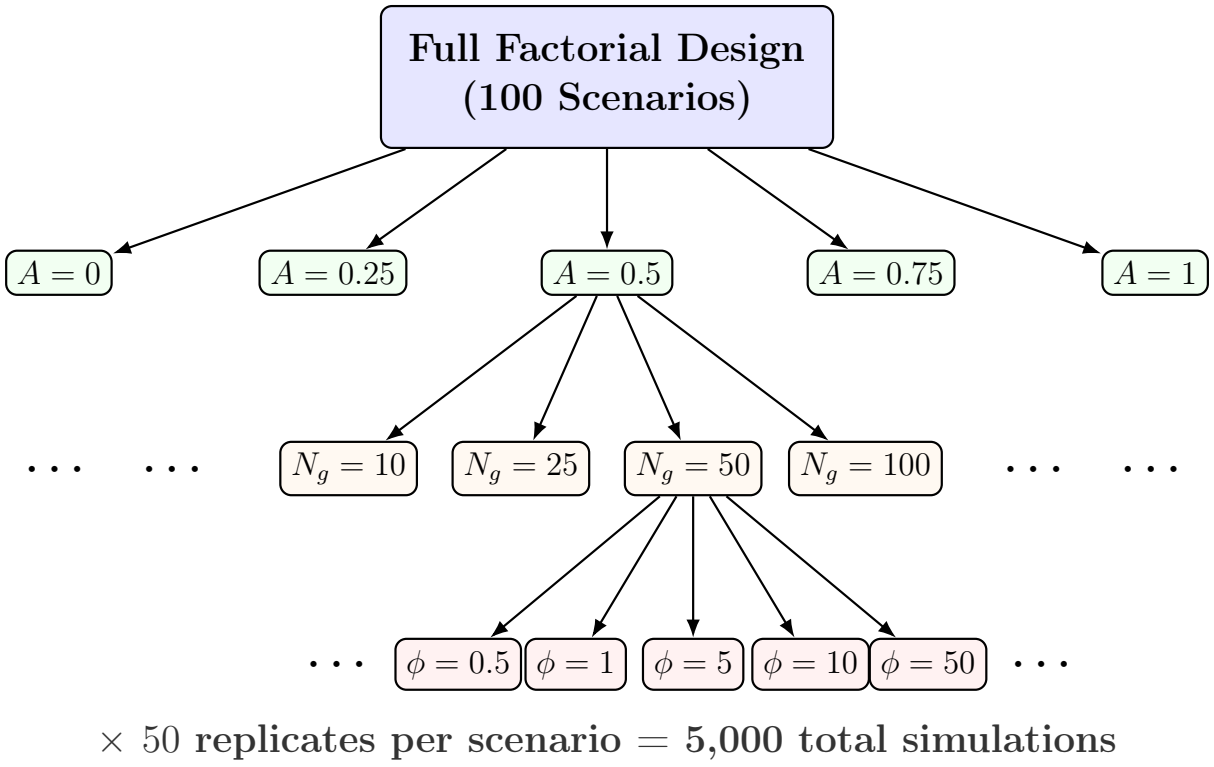


Figure 4.1: **Hierarchical representation of the full factorial simulation design.** To maintain readability, only the central branch ($A = 0.5$, $N_g = 50$) is visually expanded down to the dispersion parameter (ϕ). This combinatorial logic applies strictly across all branches, resulting in 100 unique theoretical scenarios, each independently generated 50 times, yielding a total of 5,000 specific datasets.

Note: While this full factorial design is designed to evaluate the impact of key statistical fluctuations across a broad operational range, it strategically constrains the scope of the evaluation to a fixed parameter space. Consequently, this framework addresses a delimited subset of the overarching research objectives, intentionally evaluating the pipeline's reliability strictly within the boundaries defined by these three varying factors.

4.3 Unsupervised Inference (LDA) and Latent Profile Testing (PERMANOVA)

4.3.1 Latent Structure Inference via LDA

As the LDA model inherently processes discrete data collections (Blei et al. 2003), the biological count matrix is directly formatted as a Document-Term Matrix (DTM) in accordance with the implementation standards of the `topicmodels` package (Grün and Hornik 2011). Each row represents a sample (the document) and each column represents a specific taxon (the word). This matrix conversion operationalizes the exchangeability assumption, or "bag-of-words" hypothesis, fundamental to LDA (Blei et al. 2003). Consequently, the original sequencing order of the reads is ignored in favor of a raw abundance vector per sample.

The estimation of the latent structures is performed via the `LDA()` function of the `topicmodels` package (Grün and Hornik 2011). The LDA model requires the number of latent sub-communities (K) to be fixed *a priori*. Within the strict framework of this simulation, this hyperparameter is set to strictly match the generated ground truth ($K = 5$). (It should be noted that in empirical exploratory studies, inferring the optimal K from the data constitutes an independent methodological challenge).

During the inference phase, the hyperparameters of the LDA model are not statically predefined but are dynamically estimated from the count data (`estimate.alpha = TRUE` and `estimate.beta = TRUE`) (Grün and Hornik 2011). Regarding topic prevalence, the standard LDA model assumes a global, symmetric Dirichlet prior. This formulation contrasts with the underlying generative process, where prevalence is asymmetric and directly conditioned by covariates (*cf.* Section 4.1.1.2). This symmetric parameter is iteratively optimized during inference to maximize the marginal likelihood. This specific discrepancy is intentional, as it evaluates the reliability of the unsupervised model in inferring the underlying latent structure despite a mismatched prior assumption. Furthermore, no fixed prior is imposed on the topic composition distribution; the complete matrix of taxon probabilities is directly estimated during the maximization step (Blei et al. 2003).

Regarding the inference engine, the VEM algorithm is favored over Gibbs sampling. Although Gibbs sampling offers a highly reliable stochastic approach, the VEM algorithm presents a crucial advantage in terms of computational efficiency, an indispensable requirement for the iterative processing of the full factorial design. This deterministic approach, based on Variational Bayes inference, has been successfully employed in recent large-scale microbiota studies to effectively maximize the variational lower bound of the data (Hosoda et al. 2020).

Following the convergence of the VEM algorithm, the `posterior()` function is utilized to extract the expected values of the posterior distributions, conditional on the observed data, in their natural compositional scale. This operational step yields the two fundamental inferred matrices: the topic prevalence matrix (Γ) of dimension $N \times K$, and the topic composition matrix (β) of dimension $K \times F$. Because these inferred outputs match the exact dimensions of the simulated ground truth matrices (G and B), direct mathematical comparison is enabled for the subsequent evaluation phase.

4.3.2 Resolution of Label Switching

Because the LDA model is an unsupervised mixture model, its likelihood function is invariant to the permutation of the topic components. Consequently, the assignment of topic index numbers during inference is arbitrary and inherently depends on the initialization of the algorithm. Thus, even in a theoretical scenario where the inferred sub-communities are mathematically equivalent to the simulated ground truth, there is no guarantee that the estimated "Topic 1" corresponds to the simulated "Topic 1". This phenomenon is known in Bayesian statistics as **label switching** (Sankaran and S. Holmes 2019).

To strictly evaluate the inference fidelity, it is imperative to establish a one-to-one correspondence between the estimated and simulated latent profiles. To automate this alignment, a two-step optimization procedure is implemented, inspired by the model alignment methodology documented by Fukuyama et al. (2021):

1. **Dissimilarity Matrix Computation (JSD):** The fundamental identity of a topic is defined by its taxonomic distribution over the F taxa of the vocabulary (the β_k vectors). The first step consists of calculating the distance between each inferred topic and each simulated topic. To compare these probability distributions, the JSD is employed. Unlike the Euclidean distance, which is severely impacted by the 'double-zero' problem in sparse ecological datasets (Anderson 2001; Thorsen et al. 2016), the JSD is naturally suited to compare relative abundance profiles as it formally measures the distance between probability distributions (Costea et al. 2018; Thorsen et al. 2016). This step generates a distance matrix of dimensions $K \times K$. This mathematical comparison requires the order of the taxa to be strictly identical between the two matrices, a condition fulfilled during the generative process by explicitly specifying the nomenclature in the ground truth matrix [3] , an order subsequently inherited by the count matrix and the DTM.

2. **Global Assignment via the Hungarian Algorithm:** Although the JSD matrix indicates the pairwise proximity between topics, a simple coupling based on local minimum values is insufficient. An optimal local assignment could monopolize a simulated topic and force a suboptimal pairing for the remaining distributions. Therefore, the objective is to identify the global alignment that minimizes the total sum of the JSD for all K pairs. This combinatorial optimization challenge, known as the linear assignment problem, is solved exactly by the Hungarian Algorithm (Kuhn 2005). The software implementation of this optimization relies on the R package `clue` (Hornik 2005).

Following this bipartite matching, the topic identification numbers in both the inferred composition matrix (β) and the inferred prevalence matrix (Γ) are reindexed to mirror the simulation nomenclature (1 to K). This exact alignment is implemented via the custom `alignment_topic()` function (`alignment_topic.R` script), enabling the formal evaluation of the inference accuracy.

[3] `colnames(true_p) <- paste0("taxa_", 1:sim_par$n_taxa)`

4.3.3 Testing Cohort Separation in the LDA-Inferred Prevalence Matrix via PERMANOVA test

Following the exact alignment of the latent space, the inferred prevalence matrix (Γ) serves as the quantitative input to test the association between the latent sub-communities and the predefined categorical covariate (the assigned cohort). To formally evaluate this structural difference, a PERMANOVA test is applied, utilizing the `adonis2` function from the `vegan` package (Oksanen et al. 2026).

The Bray-Curtis dissimilarity index is explicitly selected to compute the underlying distance matrix. Because the rows of the Γ matrix constitute continuous probability distributions (strictly summing to 1 per sample), they functionally mirror relative abundance profiles. While the Bray-Curtis metric is a standard measure historically favored for comparing community structures in ecology (Anderson 2001; Thorsen et al. 2016), it is also explicitly utilized in microbiome topic modeling literature to quantify dissimilarities between such mixed-membership compositional profiles (Symul et al. 2023; I. Holmes et al. 2012).

The proportion of multivariate variance explained by the cohort assignment (R^2) is assessed, and a pseudo-F statistic is computed. Under the null hypothesis (H_0), the multivariate spatial centroids of the predefined cohorts are assumed to be equivalent within the evaluated distance space. The strict statistical significance of this association (p-value < 0.05) is derived via stochastic permutations, directly establishing the binary decision criterion (rejecting or failing to reject H_0) required for the downstream performance evaluation of the analytical pipeline.

4.4 Evaluation Framework

4.4.1 Methodological Validation on a Representative Scenario

First, the integrity of the generative framework is directly verified by inspecting the simulated latent structures. The effective application of the predefined cohort effect on the latent prevalences (the generated G matrix) is confirmed. The statistical significance of these compositional differences between groups is rigorously evaluated utilizing non-parametric Wilcoxon tests, adjusted for multiple comparisons via the Benjamini-Hochberg False Discovery Rate (FDR) procedure.

Following the validation of the mathematical ground truth, the analytical perspective shifts to an exploratory analysis of the observable dataset (the count matrix). Initially, technical biases are assessed. Specifically, the distribution of the simulated sequencing depths across the predefined cohorts is evaluated using a Kruskal-Wallis test. This critical control guarantees the absence of significant technical confounding factors, ensuring that subsequent data structures are not driven by library size variations.

Subsequently, the observable relative abundance matrix is subjected to a Principal Coordinate Analysis (PCoA) based on Bray-Curtis dissimilarities. This multivariate visualization, coupled with a baseline PERMANOVA test on the observed counts, is performed to confirm that the simulated biological signal successfully is preserved despite the overdispersion noise inherent to the generative process, thereby remaining theoretically detectable prior to any unsupervised latent modeling.

Finally, the complete analytical pipeline (*cf.* Section 4.3) is applied to this isolated scenario. The unsupervised LDA inference is executed, and the correspondence between the inferred latent matrices (β and Γ) and the ground truth latent matrices (B and G) is qualitatively assessed through visual inspections (heatmaps and barplots). Quantitatively, specific performance metrics are calculated to evaluate the fidelity of the inferred outputs against the simulated ground truth. Following the topic alignment procedure, the global performance scores for topic composition (β) and topic prevalence (Γ) are defined as the mean of the diagonal elements from the $1 - JSD$ similarity matrix and the Spearman correlation matrix, respectively. Furthermore, the overall precision of the model is evaluated by reconstructing the taxon observation probability matrix ($P_{\text{LDA}} = \Gamma_{\text{LDA}} \times \beta_{\text{LDA}}$). The reconstruction error is quantified against the simulated P_{SIM} matrix via Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N \times F} \sum_{i=1}^{N} \sum_{j=1}^{F} (P_{\text{SIM},i,j} - P_{\text{LDA},i,j})^2} \tag{4.8}$$

Concurrently, the PERMANOVA test (using Bray-Curtis distance) is applied to the inferred prevalences (Γ_{LDA}) to extract the permutation-based p-value. This comprehensive assessment of a single dataset provides a concrete illustration of the pipeline's mechanics. It serves as a transitional proof of concept, defining the practical calculation of the evaluation metrics prior to their large-scale aggregation across the 5,000 simulated scenarios.

4.4.2 Evaluation Framework of the Full Factorial Design

Following the structural validation of the generative process on a specific instance, the analytical pipeline is systematically applied across the 5,000 datasets generated by the full factorial design (*cf.* Section 4.2.2). The evaluation of the unsupervised modeling relies on two distinct axes: the quantitative fidelity of the latent inference and the statistical reliability of the downstream hypothesis testing.

4.4.2.1 Evaluation and Sensitivity of Inference Fidelity

To evaluate the capacity of the unsupervised LDA algorithm to reconstruct the simulated structures, the performance metrics ($1-$ *JSD* for composition (β), Spearman correlation for prevalence (Γ), and RMSE for taxon observation probabilities (P)) are aggregated across all factorial design conditions. The trajectories of these continuous scores are then visually explored along the three simulation parameters (biological effect size A, the size parameter ϕ, and cohort size N_g) to identify performance trends.

To systematically quantify the impact of these parameters and their interactions, a sensitivity analysis is subsequently conducted. Because the evaluated continuous metrics possess divergent mathematical properties, GLMs are specifically tailored to the distribution of each score.

- The $1 - JSD$ similarity scores, being naturally bounded between 0 and 1, are modeled using a quasi-binomial GLM.
- Similarly, the Spearman correlation scores for matrix Γ are linearly transformed via $(x + 1)/2$ to map their original $[-1, 1]$ range onto the $[0, 1]$ interval, thereby enabling the application of a quasi-binomial GLM.

- The RMSE scores, which are strictly positive and typically right-skewed, are modeled utilizing a Gamma family GLM with a logarithmic link function.

Following model fitting, an analysis of deviance based on Fisher's F-test is conducted to formally quantify the respective effects of the simulation parameters and their interactions on the reconstruction scores.

Beyond average performance trends, the stability of the inference is further assessed by testing for heteroscedasticity. Specifically, median-centered Levene's tests (implemented via the `car` package) are applied to determine whether the dispersion of the performance metrics varies significantly across the levels of each simulation parameter. This median-based approach is purposefully selected to guarantee strict robustness against non-normal data distributions. The resulting p-values are adjusted for multiple comparisons using the Benjamini-Hochberg FDR procedure.

This procedure assesses the full three-factor design, formally isolating the statistical weight of each main parameter and evaluating the significance of their interactions on the algorithmic inference capabilities.

4.4.2.2 Evaluation and Sensitivity of Statistical Reliability

Beyond the exact mathematical reconstruction of the matrices, the pipeline is evaluated on its capacity to provide correct statistical decisions regarding the structural differences between the predefined cohorts. This reliability is exclusively assessed based on the empirical p-values generated by the PERMANOVA tests (significance threshold $\alpha = 0.05$). The evaluation methodology depends on the evaluated scenario:

- **Empirical Type I Error Rate (False Positives):** to quantify the rate of spurious signal detection, the empirical False Positive Rate (FPR) is computed exclusively on the subset of control simulations strictly devoid of any injected biological effect ($A = 0$). It is calculated as the percentage of simulations inappropriately yielding a significant PERMANOVA p-value.
- **Statistical Power (True Positives):** Conversely, for all scenarios where a true cohort effect is generated ($A > 0$), the statistical power is evaluated. It is quantified as the percentage of simulations successfully yielding a significant PERMANOVA p-value, illustrating the pipeline's capacity to overcome sequencing noise and reliably identify genuine structural shifts.

To systematically isolate the impact of the simulation parameters and their interactions on these detection capabilities, a sensitivity analysis is subsequently applied. Because this performance relies on binary dependent variables (significant detection coded as 1, non-significant as 0), GLMs of the binomial family with a logit link function are implemented. Separate logistic regressions are fitted for the Type I error subset and the statistical power subset, conditionally upon the presence of sufficient variance in the simulated responses to avoid quasi-complete separation. An analysis of deviance, utilizing a Likelihood Ratio (LR) test, is conducted on these binomial models to formally assess parameter significance. Finally, the resulting model coefficients are exponentiated to extract Odd Ratios (ORs), providing an interpretable quantitative measure of how variations in noise, cohort size, and biological signal alter the probability of a successful statistical detection.

4.5 Software Implementation and Reproducibility

All computational analyses are conducted in R (version 4.4.1) using RStudio (version 2026.05.0+218, "Golden Wattle"). To ensure complete reproducibility, the simulation scripts, synthetic datasets, and downstream analysis pipelines are accessible on the associated GitHub repository, and the comprehensive computational environment, including specific package versions, is detailed in Appendix A.1.

The project architecture is organized into nine discrete files, categorized into three distinct functional modules:

- **Core Utilities and Functions:** A centralized set of scripts containing all custom-built functions to maintain a clean global environment and ensure code reusability:
  - `data_simulation.R`: Contains the generative function that simulates the latent structures and the observable count matrices according to the defined parameter lists.
  - `generate_id_matrix.R`: Includes the function dedicated to generating the metadata design matrix, assigning samples to their respective predefined cohorts.
  - `alignment_topics.R`: Handles the alignment of the inferred LDA topics with the simulated ground truth using the JSD similarity matrix and the Hungarian algorithm.
  - `visualisation_functions.R`: Aggregates all custom plotting wrappers and visualization functions utilized throughout the project.
- **In-Depth Exploratory Pipeline:**
  - `simulate_1_exp(phiX).qmd`: A standalone Quarto document specifically dedicated to the qualitative, step-by-step analysis of a representative scenario.
- **Full Factorial Design Pipeline:** The large-scale simulation and inference engine, sequentially segmented into four operational scripts:
  - `0_create_design.R`: Generates the experimental design matrix. Each scenario receives a unique identifier and is saved as an RDS object, thereby freezing the design prior to any simulation.
  - `1_run_simulations.R`: Executes the simulation engine. Reproducibility is guaranteed via a strict random seed management strategy (`set.seed(1234 + rep)`) indexed to the replicate iteration.
  - `2_LDA_inference.R`: Performs the unsupervised LDA inference, executes the topic alignment, computes the evaluation metrics, and runs the PERMANOVA tests. [3]
  - `factorial_design_results.qmd`: Aggregates all synthetic score vectors to execute the sensitivity analysis (GLMs) and generate the final visualizations.

  Note: a flow control architecture is integrated into the core simulation (2) and inference scripts (3). This idempotent mechanism detects previously generated outputs, allowing calculations to be interrupted and safely resumed without redundancy,

significantly facilitating the execution of the 5,000 simulations on local computing infrastructures.

5. Results

The following sections detail the outcomes of the experimental design, sequentially addressing the qualitative validation of the generative framework, the evaluation of the latent inference fidelity, and the statistical evaluation of the cohort effect on the PERMANOVA.

5.1 Illustrative Examples: Qualitative Evaluation of the Inference Pipeline Across Extreme overdispersion Scenarios

To illustrate the inference dynamics and the impact of overdispersion on the generative model, two extreme simulations are extracted from the experimental design (*cf.* Section 4.2.2). These illustrative examples isolate the effect of the size parameter (ϕ) by maintaining a constant biological effect amplitude ($A = 1$), an identical cohort size ($N_g = 50$), and the same random generation seed (`set.seed(1261)`), while contrasting two diametrically opposed variance conditions:

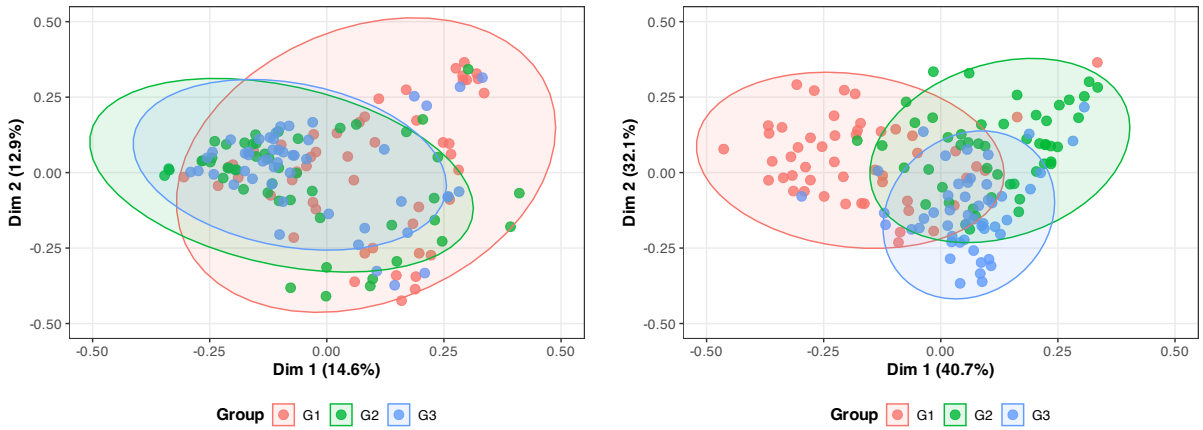
- **The High-Noise Scenario (ID: `eff1_n50_phi05_rep27`):** Characterized by severe structural overdispersion ($\phi = 0.5$), replicating the extensive biological and analytical variability of sparse microbiota composition data.
- **The Low-Noise Scenario (ID: `eff1_n50_phi50_rep27`):** Characterized by minimal overdispersion ($\phi = 50$), representing optimal analytical conditions associated with minimal stochastic perturbation and a high signal-to-noise ratio.

5.1.1 PCoA-based Assessment of Overdispersion Effects on Cohort Separation in Simulated Relative Abundances

Prior to the latent inference step, the impact of overdispersion (ϕ) on the structural visibility of the biological signal is evaluated. This assessment is performed through a visual comparison of the simulated relative abundance matrices using PCoA (Figure 5.1). This visualization demonstrates how severe stochastic noise obscures the underlying cohort structure defined in the generative ground truth.

Visual analysis of the ordinations (Figure 5.1) highlights the detrimental impact of overdispersion on the structural visibility of the inter-group biological signal. In the **High-Noise** Scenario ($\phi = 0.5$, Figure 5.1a), the underlying biological structure is largely obscured by variance. The samples exhibit high dispersion, resulting in an extensive overlap of the 95% confidence ellipses across the three groups. Mathematically, this structural dilution is reflected by a low proportion of variance explained by the primary axis (Dim 1 = 14.6%).

Conversely, in the **Low-Noise** Scenario ($\phi = 50$, Figure 5.1b), the latent structure defined during the generative process becomes distinctly observable. While the sample separation



(a) High-Noise Scenario ($\phi = 0.5$)

(b) Low-Noise Scenario ($\phi = 50$)

Figure 5.1: **Comparison of the global structure of simulated microbiota profiles across extreme overdispersion scenarios.**
PCoAs are computed from the Bray-Curtis distance matrix of simulated relative abundances. The left panel (5.1a) illustrates the High-Noise Scenario ($\phi = 0.5$), while the right panel (5.1b) represents the Low-Noise Scenario ($\phi = 50$). The biological effect amplitude ($A = 1$) and the group size ($N_g = 50$) are maintained constant across both conditions. Each point represents a single sample, colored according to its predefined group membership: G1 (red), G2 (green), and G3 (blue). The shaded ellipses outline the 95% multivariate data dispersion for each cohort.

is not absolute, the underlying cohort structure induces discernible clustering tendencies within the respective group ellipses. Furthermore, the structural information is highly centralized, with the first principal coordinate (Dim 1) capturing 40.7% of the total variance.

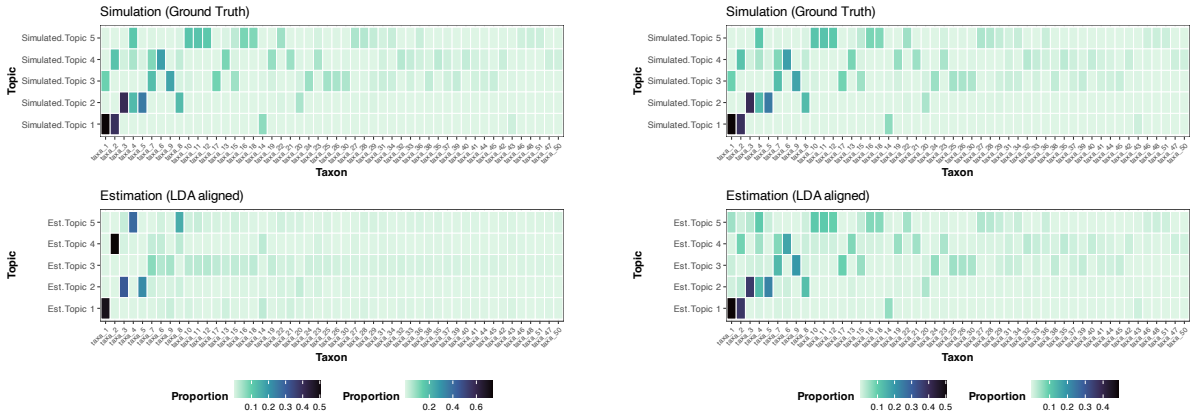
This comparison demonstrates that severe structural overdispersion (a low ϕ value) stochastically distributes the dataset variability across multiple secondary dimensions. Consequently, the effect of the cohort covariate on the relative abundances is heavily obscured prior to any processing by the latent inference algorithm (Scree plots detailing the full variance distributions are provided in Appendix A.1).

This visual structural dilution is quantified via PERMANOVA applied to the simulated relative abundances. Under minimal overdispersion ($\phi = 50$), the underlying cohort structure drives a substantial portion of the dataset variability ($R^2 = 0.301$, p-value $= 0.001$, Appendix A.2) . However, under severe structural overdispersion ($\phi = 0.5$), the variance explained by the cohort covariate collapses to 3.8% ($R^2 = 0.038$, p-value $= 0.001$, Appendix A.2). This drastic reduction confirms mathematically that prior to any inference, excessive stochastic noise fundamentally degrades the biological signal-to-noise ratio.

5.1.2 Fidelity of Latent Matrices Reconstruction Under Varying Overdispersion

This subsection evaluates the capacity of the LDA algorithm to accurately reconstruct the original latent matrices (Γ and β) and the taxon observation probability matrix (P).

Reconstruction of topic composition matrices (β): heatmap and similarity analysis The analysis relies on the comparison between the estimated composition matrices (β_{LDA}) and the simulated ground truth matrices (B_{sim}). Heatmap representations are utilized to visually identify the structural distortions induced by stochastic noise within the taxonomic profile of each topic.



(a) High-Noise Scenario ($\phi = 0.5$) (b) Low-Noise Scenario ($\phi = 50$)

Figure 5.2: **Visual comparison of the simulated and estimated taxonomic compositions of the latent topics (matrices β) across extreme overdispersion scenarios.** The heatmaps illustrate the relative proportions of the 50 most abundant taxa (x-axis) across the 5 topics (y-axis). Color intensity represents the taxon probability mass within a given topic, with darker shades indicating higher proportion. For each overdispersion condition (left panel 5.2a: High-Noise Scenario, $\phi = 0.5$; right panel 5.2b: Low-Noise Scenario, $\phi = 50$), the upper matrix displays the simulated ground truth (B_{sim}) and the lower matrix presents the corresponding estimation by the LDA model (β_{LDA}) following topic alignment. The biological effect amplitude ($A = 1$) and the cohort size ($N_g = 50$) are maintained constant.

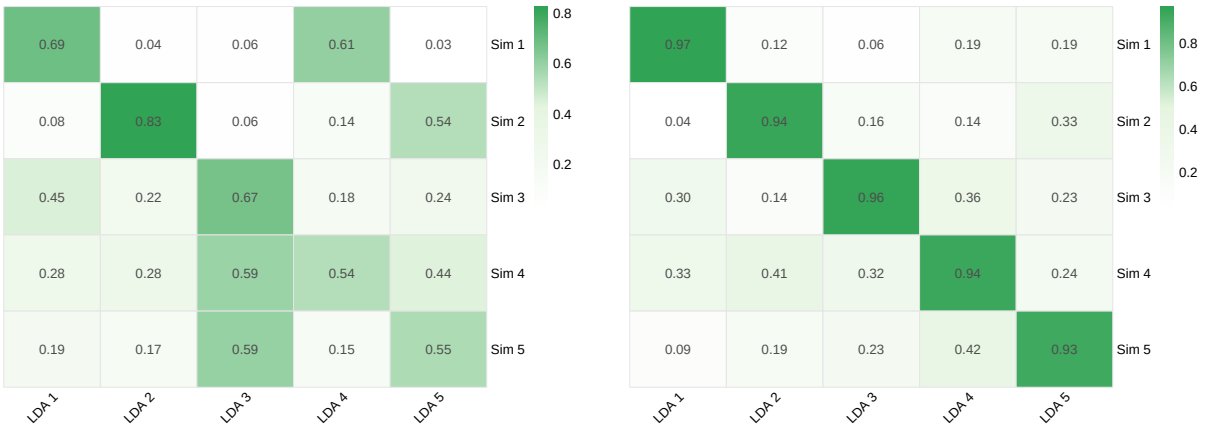
Observation of the simulated ground truth matrices (upper panels, Figure 5.2) confirms that the generative pipeline produces distinct taxonomic profiles for each topic. This baseline representation highlights highly abundant taxa strictly confined to specific topics, providing compositional signatures. These simulated structural profiles are strictly identical between the two scenarios. This invariance confirms that the size parameter (ϕ) solely modulates the final count matrix generation without altering the underlying generative probabilities.

Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$, Figure 5.2a), the inferred matrix (β_{LDA}) exhibits marked structural divergence from the ground truth. The inference process yields artificially concentrated taxonomic distributions. The probability mass is predominantly monopolized by one or two highly dominant taxa per topic (indicated by isolated dark cells), to the detriment of intermediate-abundance taxa whose compositional signal is lost.

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$, Figure 5.2b), the estimated β_{LDA} matrix reproduces the theoretical distributions with high fidelity. The latent inference accurately captures the core characteristic taxa of each topic. However, variations in probability masses remain, and certain low-abundance taxa are attenuated

in the estimated matrix, a structural loss particularly observable within the compositional profile of "Topic 3". This visual comparison indicates that severe count overdispersion degrades the inference of taxonomic compositions. The estimation process is skewed toward an overestimation of major taxa at the expense of the true sub-community richness.

These visual observations (Figure 5.2) are quantified through the inspection of similarity matrices (Figure 5.3). These matrices provide a mathematical evaluation of the reconstruction fidelity regarding the latent taxonomic compositions.



(a) High-Noise Scenario ($\phi = 0.5$)

(b) Low-Noise Scenario ($\phi = 50$)

Figure 5.3: **Similarity matrices evaluating the reconstruction fidelity of the taxonomic compositions (matrices β).**
The matrices compare the simulated ground truth topics (rows) to the topics inferred by the LDA model (columns) following the alignment step. The structural resemblance between two profiles is quantified by a similarity score computed as $1 - JSD$. A score of 0 indicates diametrically opposed compositions, while a score of 1 represents identical taxonomic probabilities. The left panel (5.3a) illustrates the High-Noise Scenario ($\phi = 0.5$), and the right panel (5.3b) represents the Low-Noise Scenario ($\phi = 50$). The main diagonal highlights the optimal matching between the simulated topics and their estimated equivalents computed via the Hungarian algorithm.

Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$, Figure 5.3a), the inference fidelity is degraded. The alignment coefficients on the main diagonal drop substantially, yielding a low average similarity score of 0.66 (mean of the diagonal). More critically, this degradation is accompanied by a loss of compositional resolution, observable off the diagonal. For instance, the inferred sub-community "LDA 3" exhibits simultaneous similarities with simulated topics 3, 4, and 5 (respective scores of 0.67, 0.59, and 0.59). These mathematical overlaps demonstrate that severe overdispersion degrades the resolving power of the algorithm, resulting in the artificial convolution of distinct taxonomic profiles into a single inferred signature.

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$, Figure 5.3b), the inferred matrix exhibits high structural fidelity. The maximum similarity coefficients are strictly confined to the main diagonal (with values ranging between 0.93 and 0.97), distinctly standing out from the off-diagonal background . This configuration indicates that the LDA algorithm isolates and reconstructs each compositional profile with high specificity, yielding a mean diagonal similarity score of 0.95. In summary, this comparative

analysis demonstrates that extreme overdispersion limits the capacity of the model to isolate specific taxonomic signatures. This limitation translates into a reduction in reconstruction fidelity and structural overlaps between the inferred latent dimensions.

5.1.2.1 Reconstruction of Sample-Level Topic Prevalence (Γ): Barplot and Spearman Correlation Analysis

Similar to the taxonomic compositions, the sample-level topic prevalences are evaluated by comparing the inferred proportions (Γ_{LDA}) to the simulated ground truth (G_{sim}). Stacked bar charts (Figure 5.4) are utilized to visually assess whether the cohort-specific compositional signatures are maintained under varying levels of structural overdispersion.

This evaluation connects the inferred latent structure to the cohort covariate. Observation of the simulated ground truth matrices (left panels, Figure 5.4) confirms the implementation of the generative experimental design (*cf.* Table 4.1) : a distinct over-representation of Topic 1 in Group 1, a dominance of Topic 2 in Group 2, and a balanced distribution of Topics 3, 4, and 5 in Group 3 are visually observable. In accordance with the generative architecture, these theoretical baseline profiles remain strictly invariant across overdispersion conditions.

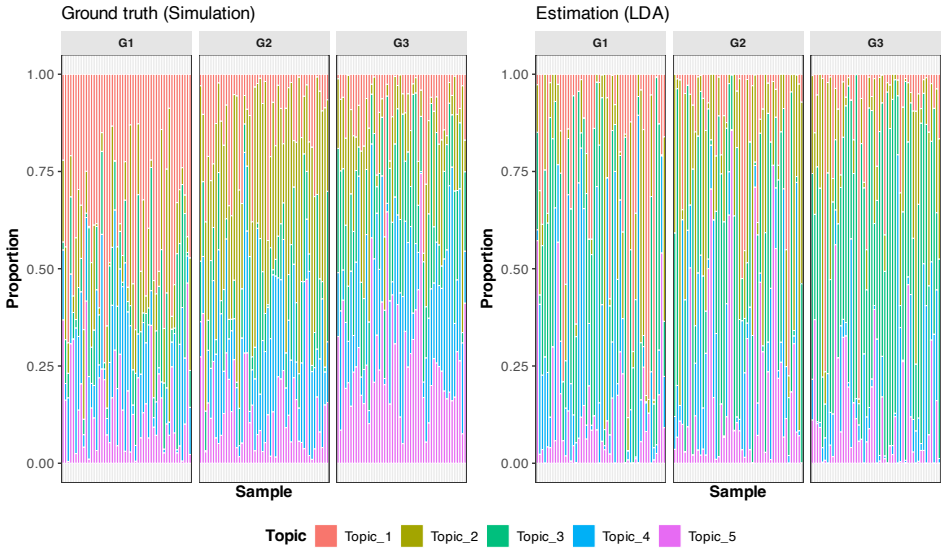
Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$, Figure 5.4a), the inference process results in a loss of the original cohort structure. The inferred matrix (Γ_{LDA}, right panel) exhibits highly variable and smoothed profiles, obscuring the predefined group signatures. The expected dominance of Topic 1 and Topic 2 within their respective cohorts seems attenuated. Simultaneously, Group 3 exhibits an artificial over-representation of Topic 3, diverging significantly from the generative baseline. This major visual discrepancy indicates that extreme count variance structurally degrades the capacity of the latent inference to recover the underlying cohort-specific biological effects.

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$, Figure 5.4b), the estimated Γ_{LDA} matrix more faithfully reproduces the structural ground truth. The latent inference accurately tends to recover the expected topic proportions and to preserve the signature of each predefined cohort, confirming higher reconstruction fidelity under optimal variance conditions.

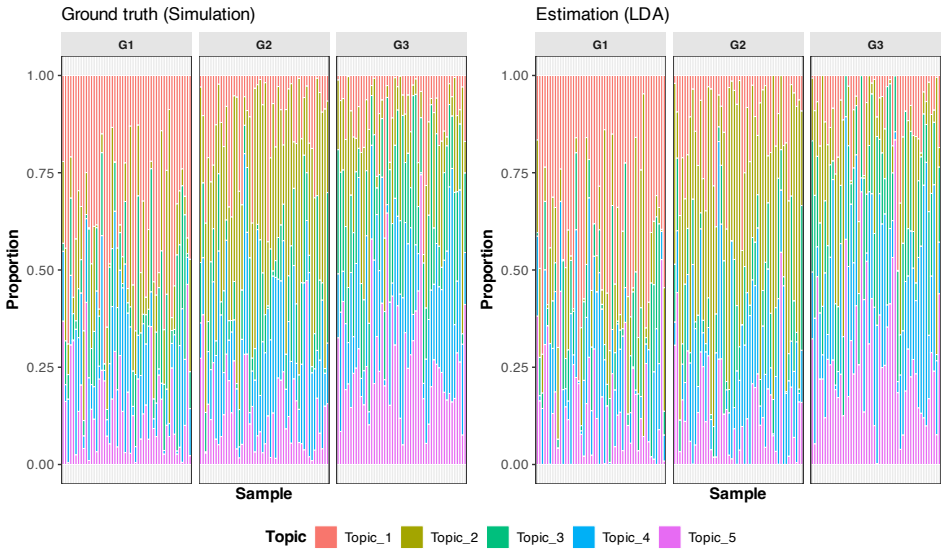
Inspection of the Spearman correlation matrices (Figure 5.5) quantitatively confirms the visual discrepancies observed in the prevalence profiles of the Γ matrix.

Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$, Figure 5.5a), the inference fidelity of the sample prevalence is degraded. The alignment coefficients on the main diagonal drop substantially, reducing the mean correlation score to 0.30 (mean of the diagonal), a structural loss further confirmed by an equivalent Mantel correlation score of 0.30 (Appendix A.3). More critically, this degradation is accompanied by a severe loss of distributional resolution, observable off the diagonal. For instance, the inferred prevalence of "LDA 3" exhibits high simultaneous correlations with simulated topics 3, 4, and 5 (respective scores of 0.49, 0.20, and 0.45). These mathematical overlaps demonstrate that excessive stochastic noise degrades the capacity of the model to isolate distinct prevalence distributions across the samples.

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$, Figure 5.5b), the inferred matrix exhibits high structural fidelity. The maximum correlation coefficients

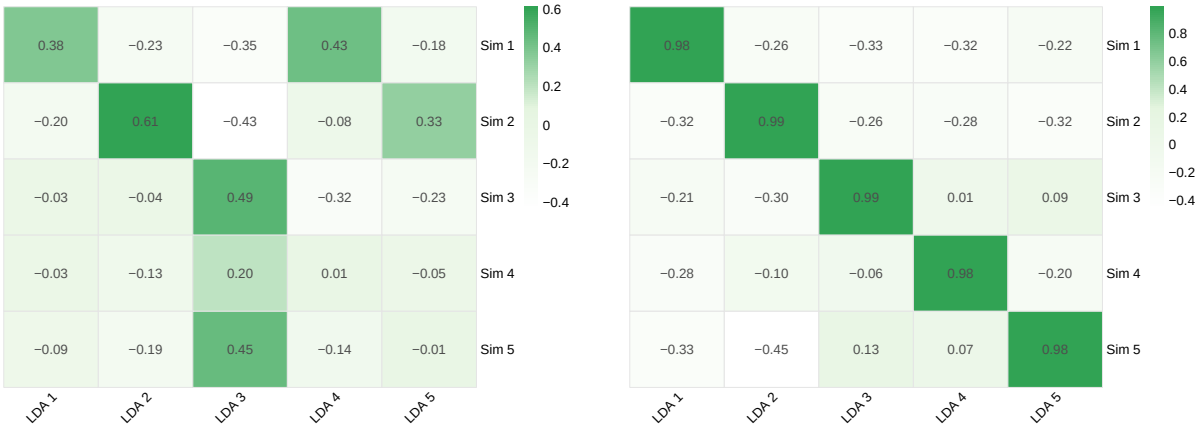


(a) High-Noise Scenario ($\phi = 0.5$)



(b) Low-Noise Scenario ($\phi = 50$)

Figure 5.4: **Visual comparison of the sample-level topic prevalence (matrices Γ) across extreme overdispersion scenarios.**
The stacked bar charts represent the relative probability mass (y-axis) of each latent topic (red = T1, khaki = T2, green = T3, blue = T4, and purple = T5) within the individual samples (x-axis) of the simulated cohorts. The upper graph (5.4a) illustrates the High-Noise Scenario ($\phi = 0.5$) and the lower graph (5.4b) represents the Low-Noise Scenario ($\phi = 50$). The biological effect amplitude ($A = 1$) and the cohort size ($N_g = 50$) are maintained constant. For each condition, the left panel displays the simulated generative prevalences (G_{sim}), while the right panel presents the corresponding proportions inferred by the LDA model (Γ_{LDA}) following alignment. Samples are separated into three distinct facets according to their predefined cohort affiliation (G1, G2, G3).



(a) High-Noise Scenario ($\phi = 0.5$) (b) Low-Noise Scenario ($\phi = 50$)

Figure 5.5: **Spearman correlation matrices evaluating the reconstruction fidelity of the sample-level topic prevalences (matrices Γ).**
The matrices compare the ground truth generative prevalences (rows) to the prevalences inferred by the LDA model (columns) following the alignment step. The structural matching is quantified by the Spearman rank correlation coefficient, where a value of 1 indicates complete preservation of the sample ranking according to topic probability mass, 0 indicates no monotonic relationship, and -1 corresponds to a fully inverse ranking. The left panel (5.5a) illustrates the High-Noise Scenario ($\phi = 0.5$) and the right panel (5.5b) represents the Low-Noise Scenario ($\phi = 50$). The main diagonal highlights the optimal matching between the simulated topics and their estimated equivalents computed via the Hungarian algorithm.

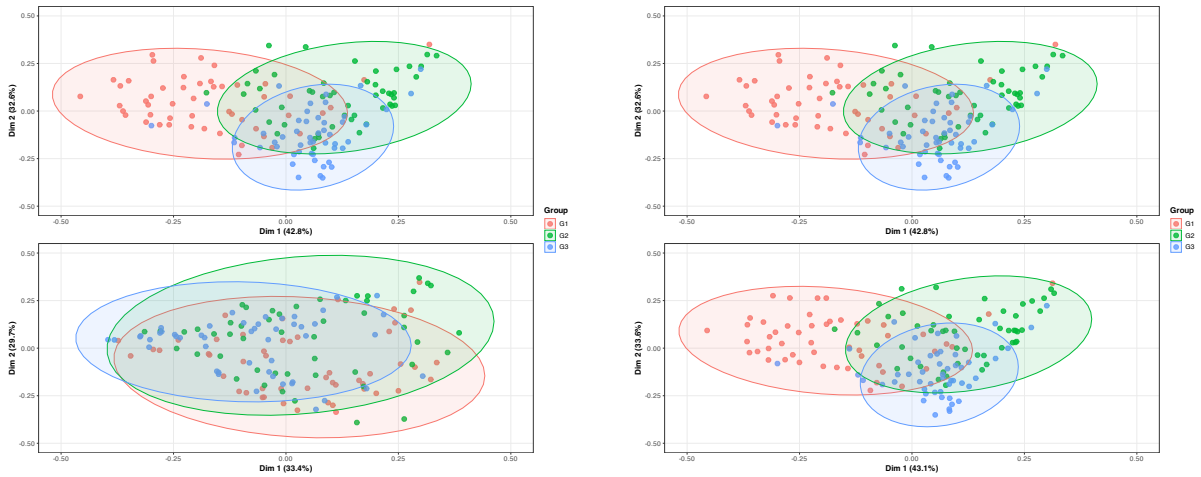
are strictly confined to the main diagonal (with values ranging between 0.98 and 0.99), distinctly standing out from the off-diagonal background noise. This configuration confirms that the LDA algorithm isolates and reconstructs the latent profile of each sample with high specificity, yielding a mean diagonal correlation score of 0.986 and an equivalent Mantel test score of 0.98 (Appendix A.3).

In summary, this comparative analysis formally demonstrates that severe structural overdispersion fundamentally degrades the capacity of the inference algorithm to accurately rank samples based on their latent topic abundance. This limitation destroys the cohort-specific compositional signatures, critically compromising the biological relevance of the inferred Γ matrix.

5.1.2.2 Restoration of the Reconstruction of Taxa Observation Probabilities (P): PCoA and RMSE Evaluation

The global reconstruction fidelity is evaluated through the analysis of the noise-free taxon observation probability matrices (P). This assessment is conducted visually using PCoA and quantitatively supported by calculating the RMSE and computing Mantel correlation statistics between P_{sim} and P_{LDA}.

Visual analysis of the ordinations (Figure 5.6) highlights the detrimental impact of structural overdispersion on the capacity of the LDA algorithm to reconstruct the underlying biological signal (P). Observation of the simulated ground truth matrices (upper panels) confirms the stability of the generative pipeline: the baseline ordinations are strictly iden-



(a) High-Noise Scenario ($\phi = 0.5$)

(b) Low-Noise Scenario ($\phi = 50$)

Figure 5.6: **Visual comparison of the taxon observation probabilities (P) across extreme dispersion scenarios.** PCoA is computed from the Bray-Curtis distance matrix derived from the underlying observation probabilities ($P = \Gamma \times \beta$). The left panel (5.6a) illustrates the High-Noise Scenario ($\phi = 0.5$), the right panel (5.6b) represents the Low-Noise Scenario ($\phi = 50$). The biological effect amplitude ($A = 1$) and the cohort size ($N_g = 50$) are maintained constant. For each condition, the upper graph presents the simulated ground truth matrix (P_{sim}) and the lower graph displays the corresponding matrix inferred by the LDA model (P_{LDA}). Each point corresponds to the structural arrangement of a single sample, colored according to its predefined cohort membership: Group 1 (G1) in red, Group 2 (G2) in green, and Group 3 (G3) in blue. The shaded ellipses outline the 95% multivariate data dispersion for each cohort.

tical across both scenarios. This invariance demonstrates that the generative inter-group variance is unaffected by the size parameter (ϕ), which mathematically intervenes strictly downstream during the generation of the final count matrices.

Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$, Figure 5.6a), the inferred biological structure (P_{LDA}) deviates significantly from the generative baseline. The inferred sample profiles exhibit high dispersion, resulting in a pronounced overlap of the 95% confidence ellipses across the three groups. Mathematically, this structural dilution is characterized by a decrease in the variance proportion captured by the first principal coordinate, dropping to 33.4% (compared to 42.8% for the ground truth). This degraded reconstruction is quantitatively confirmed by an elevated RMSE of 0.0214 and a low Mantel correlation score of 0.685 (Appendix A.4).

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$, Figure 5.6b), the estimated latent structure reproduces the simulated matrix with high fidelity. The original cohort differentiation is accurately restored, with the samples forming distinct clusters within their respective bounding ellipses. Furthermore, the structural variance is better preserved along the primary axis, with the first inferred dimension (Dim 1) capturing 43.1% of the total variance, closely matching the simulated baseline (42.8%). This optimal reconstruction is quantitatively supported by a high Mantel score of 0.995 and an low RMSE equal to 0.003 (Appendix A.4).

In summary, this comparative analysis demonstrates that severe stochastic overdispersion

fundamentally limits the structural denoising capacity of the latent inference model. The algorithm fails to entirely resolve the introduced variance scatter, yielding a degraded reconstruction that partially obscures the underlying cohort effect within the inferred probability matrix (Scree plots detailing the variance distributions are provided in Appendix A.2).

5.1.2.3 PERMANOVA-Based Quantification of Cohort Signal Preservation (Matrix Γ)

This qualitative evaluation concludes with the statistical verification of cohort signal preservation within the inferred prevalence profiles (matrices Γ). PERMANOVA tests are performed to quantify the proportion of variance (R^2) explained by the "cohort" covariate. This procedure evaluates the extent to which the underlying cohort differences are maintained following the latent inference process, particularly under high variance conditions.

Analysis of the simulated ground truth matrices (G_{sim}) provides a validation of the generative pipeline architecture. Regardless of the dispersion parameter (ϕ), the PERMANOVA results on the generative prevalences remain strictly identical ($R^2 = 0.283$, p-value $= 0.001$) (Appendix A.5). This constancy corroborates previous visual observations (Figure 5.4), confirming that stochastic overdispersion exclusively acts upon the final count generation and does not alter the underlying generative topic distributions.

Under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$), the inferred matrix (Γ_{LDA}) exhibits a degradation of the cohort signal. Although the PERMANOVA test applied to the inferred prevalences still yields a statistically significant group effect (p-value $= 0.001$), the explanatory power of the cohort covariate collapses. The coefficient of determination (R^2) drops to 0.066 (Appendix A.5), indicating that only 6.6% of the variance in the estimated profiles is attributable to the cohort affiliation, compared to 28.3% in the generative ground truth, a more than fourfold reduction.

Conversely, under minimal overdispersion (Low-Noise Scenario, $\phi = 50$), the matrix inferred by the LDA model preserves the cohort signal with fidelity. The PERMANOVA test reveals a significant group effect (p-value $= 0.001$) coupled with an explained variance proportion ($R^2 = 0.284$, Appendix A.5) that is nearly identical to the simulated baseline. Consequently, the biological structuring of the cohort is accurately restored and analytically exploitable.

In conclusion, this reduction in explained variance quantitatively synthesizes the structural degradations observed in previous analyses. It demonstrates that extreme variance smooths the inferred prevalence profiles and dilutes inter-group boundaries. While statistical significance is retained on the heavily dispersed estimates, the severe loss of structural information fundamentally limits the capacity of the inferred matrix to effectively discriminate sample cohort affiliations.

5.2 Systematic Evaluation of Inference Fidelity and Statistical Power

This section transitions from the qualitative analysis of extreme scenarios to a comprehensive, quantitative evaluation of the latent inference operational limits. To assess the capacity of the LDA algorithm to infer underlying structures in the presence of a predefined cohort covariate, inference fidelity and statistical power are systematically quantified across the full factorial experimental design (*cf.* in Section 4.2.2).

The simulated dataset encompasses the systematic variation of three critical generative parameters: the biological effect amplitude ($A \in \{0, 0.25, 0.5, 0.75, 1\}$), the cohort size, defined as the number of samples per group ($N_g \in \{10, 25, 50, 100\}$), and the degree of overdispersion, governed by the size parameter (ϕ) of the negative binomial distribution ($\phi \in \{0.5, 1, 5, 10, 50\}$).

This subsection details the evolution of three global evaluation metrics calculated across the full experimental design. The objective is to define the boundaries of stochastic overdispersion and sample size within which the latent data matrices are accurately reconstructed. Each metric is supplemented by a sensitivity analysis (analysis of deviance) to identify the individual factors and interactions that govern the inference fidelity.

5.2.1 Topic composition reconstruction (matrix β): JSD similarity scores

The evaluation of the β matrix reconstruction fidelity begins with the visual analysis of the similarity scores ($1 - JSD$) across all simulated parameter combinations (Figure 5.7). This metric quantifies the precision of the inferred taxonomic composition for each latent topic.

The reconstruction scores for the β matrices range between 0.50 and 0.95 (Figure 5.7). Trajectory analysis indicates that the dispersion parameter (ϕ) constitutes the primary factor influencing inference fidelity. Independently of other parameters, a decrease in ϕ (increased overdispersion) induces a reduction in the mean scores. Conversely, as overdispersion attenuates, the similarity scores improve monotonically, reaching their maxima at $\phi = 50$. A distinct discontinuity in reconstruction fidelity is observed between the $\phi = 1$ and $\phi = 5$ thresholds.

The magnitude of the biological effect (A, x-axis) exhibits a positive, albeit secondary, impact. The mean trend lines display a slight upward slope as A increases. This dynamic suggests that a pronounced biological signal between the cohorts marginally improves the structural resolution of the latent inference process.

The cohort size (N_g, facets) maintains the general trajectory of the curves while inducing specific interaction effects. Mean scores increase slightly and consistently as N_g expands. Furthermore, an interaction between sample size and overdispersion modulates the variance of the estimates. In restricted cohorts ($N_g = 10$ and 25), point dispersion remains relatively homogeneous across noise conditions. In expanded cohorts ($N_g = 100$), this variance behavior diverges: highly overdispersed scenarios ($\phi \leq 1$) exhibit increased estimate dispersion, whereas low-noise scenarios ($\phi \geq 5$) form highly clustered estimations. This latter configuration creates a high-precision plateau tightly constrained around a score

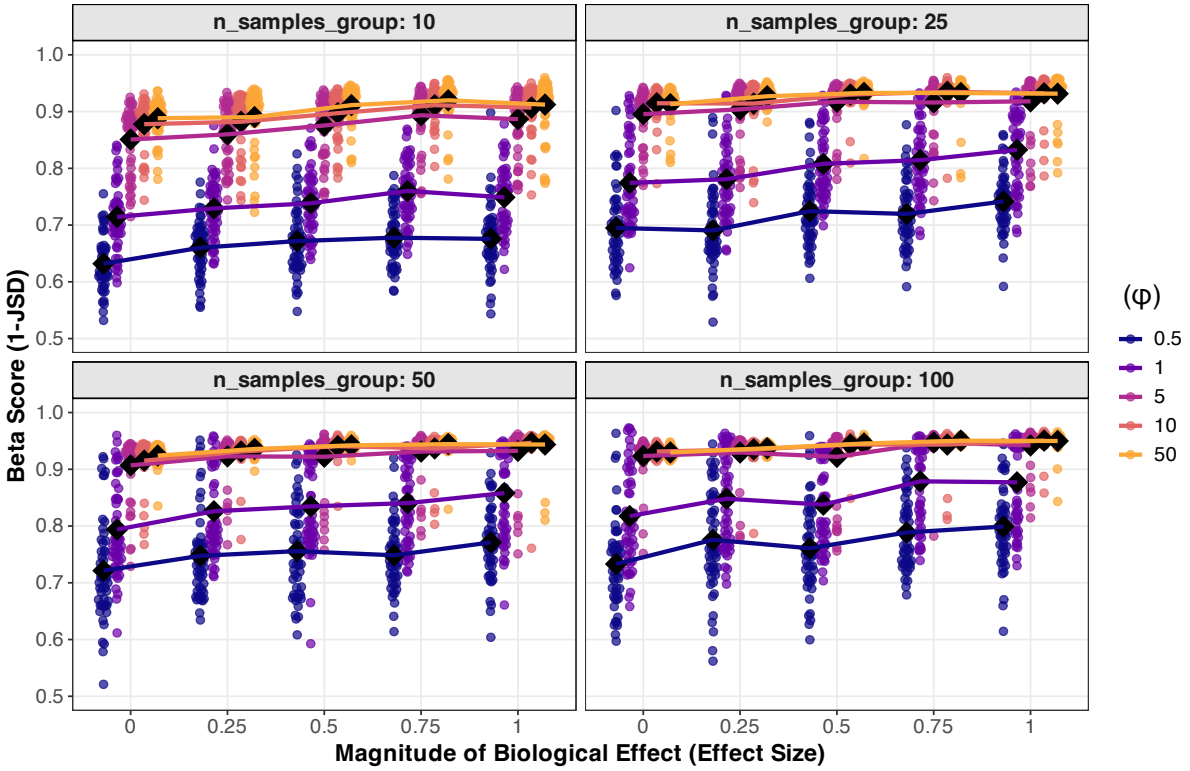


Figure 5.7: **Evolution of the taxonomic composition reconstruction scores (matrix β) across the experimental design.** The graph illustrates the distribution of the similarity scores $(1 - JSD)$ computed between the simulated generative matrices (B_{sim}) and the inferred matrices (β_{LDA}) following alignment. A score of 1 indicates perfect structural reconstruction of the taxonomic composition. A score of 0 indicates complete divergence. Each point represents an independent simulation. The data are stratified into facets according to the cohort size (N_g, ranging from 10 to 100 samples per group). The x-axis represents the magnitude of the imposed biological effect (A). Colors indicate the value of the size parameter ϕ (ranging from blue for high overdispersion (O.5) to orange for low overdispersion (50)). Black diamonds denote the mean score for each condition. Trend lines are colored according to ϕ.

of 0.95. These observations suggest a non-linear effect of the dispersion parameter on estimate variance, which is strictly modulated by sample size.

The preceding visual observations (Figure 5.7) are quantified through an analysis of deviance, summarized in Table 5.1. Given that the similarity scores $(1 - JSD)$ are strictly bounded within the $[0, 1]$ interval, this sensitivity analysis is derived from a GLM utilizing a quasibinomial family. This statistical evaluation confirms the structural relevance of the experimental design, as all primary main effects demonstrate high statistical significance (A full version of the table is provided in Appendix A.6).

Analysis of the F-statistics confirms that the dispersion parameter (ϕ) is the dominant factor governing reconstruction fidelity ($F_{4,4900} = 2975.32$, $p < 0.001$). Cohort size (N_g) is identified as the secondary influencing factor ($F_{3,4900} = 499.06$, $p < 0.001$). As visually observed, the biological effect amplitude (A) exerts a statistically significant, albeit lesser, impact on the structural resolution of the latent inference ($F_{4,4900} = 82.24$, $p < 0.001$).

Table 5.1: **Sensitivity analysis of the taxonomic composition reconstruction scores (matrix β) – Analysis of deviance.** The table presents the F-statistics and associated empirical p-values for the main effects and interactions evaluated across the experimental design ($N = 5000$ simulations). Degrees of freedom for the numerator and denominator (Residuals: 4900) are indicated.

Factor / Interaction	F**-value**	p**-value**
Dispersion Parameter (ϕ)	2975.32	< 0.001
Cohort Size (N_g)	499.06	< 0.001
Biological Effect Amplitude (A)	82.24	< 0.001
$\phi \times N_g$	2.37	0.005
$N_g \times A$	2.99	< 0.001
$\phi \times A$	1.39	0.134
$\phi \times N_g \times A$	0.53	0.967

Examination of the interaction terms highlights a significant modulation between stochastic overdispersion and cohort size ($\phi \times N_g$) ($F_{12,4900} = 2.37$, $p = 0.005$). This statistical result formalizes the previously observed variance amplification: expanded cohort sizes exacerbate the structural impact of extreme overdispersion, inducing high estimate variance under noisy conditions, while simultaneously converging toward a highly precise plateau under optimal variance conditions.

Regarding interactions involving the biological effect amplitude (A), the interaction with cohort size ($N_g \times A$) is statistically significant ($F_{12,4900} = 2.99$, $p < 0.001$). Conversely, the interaction between overdispersion and biological effect ($\phi \times A$) does not reach the significance threshold ($F_{16,4900} = 1.39$, $p = 0.134$). Finally, the absence of a significant three-way interaction ($F_{48,4900} = 0.53$, $p = 0.967$) highlights the stability of the inference process: the reconstruction fidelity is fundamentally constrained by the joint relationship between stochastic noise and sample size, largely independent of the underlying biological signal amplitude.

Beyond the evaluation of mean reconstruction scores, the dispersion of the fidelity estimates is formally assessed to determine if the inference uncertainty is subject to heteroskedasticity. Levene's test is conducted to quantify the influence of generative parameters on the variance of the similarity scores. The results (Table 5.2) confirm that the reconstruction variance is non-homogeneous across all tested experimental factors. The size parameter (ϕ) and the cohort size (N_g) exert a impact on the dispersion of the estimates, as evidenced by extremely significant p-values. The biological effect amplitude (A) also contributes significantly to this heteroskedasticity. These statistical findings corroborate the visual patterns (Figure 5.7), where restricted cohort sizes and high stochastic overdispersion manifest as increased estimate spread, whereas expanded cohorts and optimal dispersion parameters converge toward a high-precision fidelity plateau.

Table 5.2: **Impact of simulation parameters on the variance of compositional fidelity scores (β).** Results of Levene's tests evaluating the homogeneity of variance across experimental factors. p-values are adjusted using the Benjamini-Hochberg FDR method to account for multiple testing.

Factor	**Raw p-value**	**Adjusted p-value (FDR)**
Biological Effect Size (A)	3.86×10^{-3}	3.86×10^{-3}
Size Parameter (ϕ)	7.06×10^{-246}	2.12×10^{-245}
Cohort Size (N_g)	1.90×10^{-27}	2.85×10^{-27}

5.2.2 Sample-level topic prevalence reconstruction (matrix Γ): Spearman correlation scores

The inference fidelity of the sample-level topic prevalences (matrix Γ) is evaluated using the Spearman rank correlation coefficient. The reported metric corresponds to the mean diagonal correlation score computed between the generative prevalences (G_{sim}) and the inferred proportions (Γ_{LDA}). This metric formally quantifies the preservation of sample rankings with respect to topic probability mass across the experimental design (Figure 5.8).

Global observation of the trajectories (Figure 5.8) reveals inference dynamics strictly analogous to those observed for the β matrix. The dispersion parameter (ϕ) remains the dominant degradation factor. Furthermore, the variance amplification effect induced by cohort size (N_g) under highly overdispersed conditions is consistently preserved. Consequently, this analysis focuses on the structural distinctions between the inference of taxonomic compositions and the inference of topic prevalences.

The primary distinction lies in the absolute range of the inferred scores. While the evaluation of the β matrix plateaued within an interval of 0.50 to 0.95, the correlation scores for the Γ matrix span a broader spectrum. Scores range from near zero (indicating near-random rank preservation at $\phi = 0.5$) to a maximum of 0.994 (indicating optimal structural fidelity at $\phi = 50$). This expanded dynamic is partially attributable to the distinct mathematical boundaries of the specific metrics employed (Spearman correlation versus Jensen-Shannon divergence).

The second, structurally more significant distinction pertains to the convergence toward the optimal fidelity plateau. The latent inference process stabilizes the macroscopic sample-level prevalences more rapidly than the internal taxonomic compositions. The high-precision plateau (scores approaching 1 for low-noise scenarios) is consistently reached within restricted cohort sizes ($N_g = 50$). In contrast, the β matrix required expanded cohorts ($N_g = 100$) to achieve equivalent stabilization around 0.95. This observation indicates that the structural inference of macroscopic sample profiles requires a lower quantitative information threshold compared to the precise resolution of fine-scale taxonomic signatures.

The sensitivity analysis (analysis of deviance, Table 5.3) confirms the relative impact of the main effects previously identified for the reconstruction of the taxonomic compositions (β), while highlighting interaction dynamics specific to the inference of sample-level prevalences. To satisfy the domain requirements of the variance model, the Spearman correlation scores

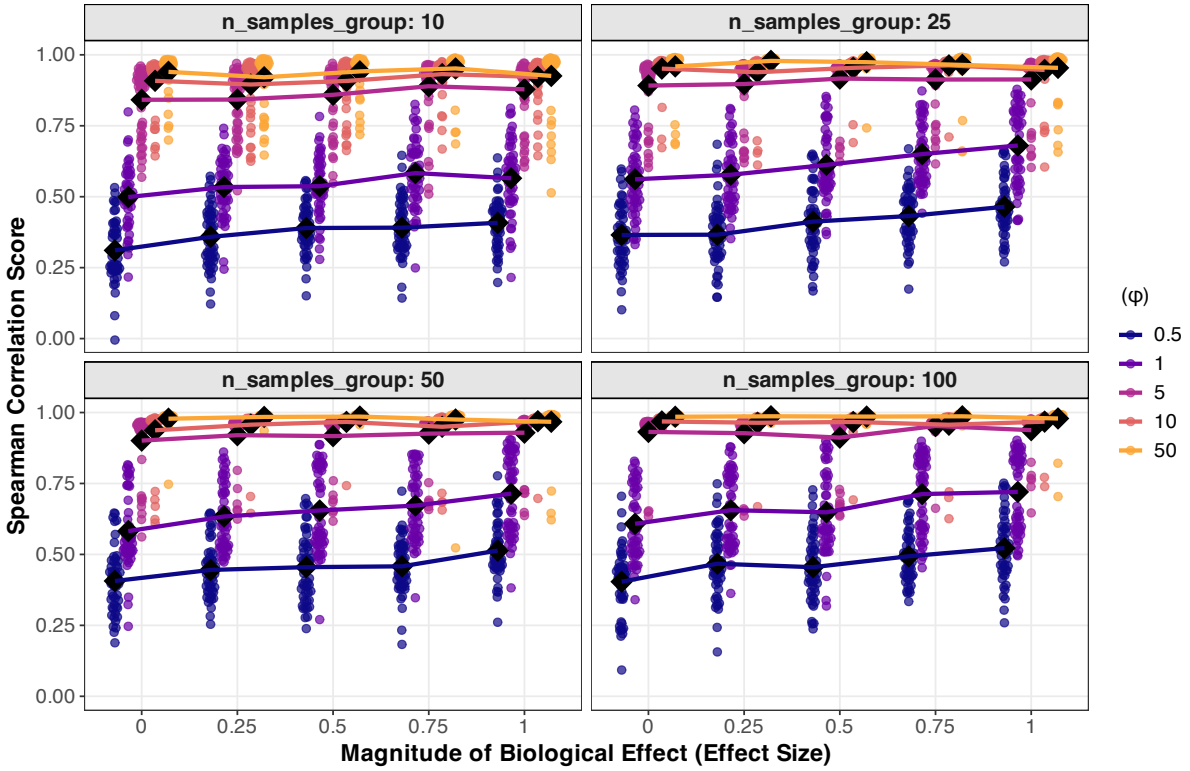


Figure 5.8: **Evolution of the sample-level topic prevalence reconstruction scores (matrix Γ) across the experimental design.** The graph illustrates the distribution of the mean diagonal Spearman correlation scores computed between the generative baseline matrices (G_{sim}) and the inferred matrices (Γ_{LDA}) following alignment. A score of 1 indicates exact structural preservation of sample rankings regarding topic proportions, whereas a score of 0 indicates a random distribution and a score of -1 indicates a perfectly inverse ranking. Each point represents an independent simulation. The data are stratified into facets according to the cohort size (N_g, ranging from 10 to 100 samples per group). The x-axis represents the magnitude of the imposed biological effect (A). Colors indicate the dispersion parameter ϕ (ranging from blue for high overdispersion to orange for low overdispersion). Black diamonds denote the mean score for each condition. Trend lines are colored according to ϕ.

(originally bounded between -1 and 1) are linearly transformed to the $[0, 1]$ interval via the function $f(x) = \frac{x+1}{2}$. Subsequently, GLM specifying a quasibinomial family is fitted to compute the analysis of deviance (A full version of the table is provided in Appendix A.7).

The dispersion parameter (ϕ) remains the dominant factor governing the majority of the variance in the prevalence correlation scores ($F_{4,4900} = 4247.43$, $p < 0.001$). Cohort size N_g ($F_{3,4900} = 122.97$, $p < 0.001$) and biological effect amplitude A ($F_{4,4900} = 30.45$, $p < 0.001$) retain the second and third positions, respectively, in terms of structural influence.

Regarding the structure of the two-way interactions, a specific dynamic emerges compared to the taxonomic reconstructions. The interaction between stochastic overdispersion and cohort size ($\phi \times N_g$) remains the most impactful interaction ($F_{12,4900} = 12.72$, $p < 0.001$). Contrary to the β matrix inference, the interaction between overdispersion and the biological effect ($\phi \times A$) reaches statistical significance ($F_{16,4900} = 2.03$, $p = 0.009$). This

Table 5.3: **Sensitivity analysis of the sample-level prevalence reconstruction scores (matrix Γ) – Analysis of deviance.** The table presents the F-statistics and associated empirical p-values for the main effects and interactions evaluated across the experimental design ($N = 5000$ simulations). Degrees of freedom for the numerator and denominator (Residuals: 4900) are indicated.

Factor / Interaction	F**-value**	p**-value**
Dispersion Parameter (ϕ)	4247.43	< 0.001
Cohort Size (N_g)	122.97	< 0.001
Biological Effect Amplitude (A)	30.45	< 0.001
$\phi \times N_g$	12.72	< 0.001
$\phi \times A$	2.03	0.009
$N_g \times A$	1.06	0.392
$\phi \times N_g \times A$	0.91	0.652

Table 5.4: **Impact of simulation parameters on the variance of prevalence fidelity scores (Γ).** Results of Levene's tests evaluating the homogeneity of variance across experimental generative factors. Empirical p-values are adjusted using the False Discovery Rate (FDR) method to account for multiple comparisons.

Factor	**Raw** p**-value**	**Adjusted** p**-value (FDR)**
Biological Effect Size (A)	5.20×10^{-5}	5.20×10^{-5}
Size Parameter (ϕ)	2.83×10^{-209}	8.48×10^{-209}
Cohort Size (N_g)	2.70×10^{-8}	4.06×10^{-8}

statistical formalization indicates that the accurate structural attribution of latent profiles to individual samples is more sensitive to the combined effects of sampling effort, structural variance, and the underlying biological signal amplitude.

Finally, the two-way interaction between cohort size and biological effect ($N_g \times A$) is not statistically significant ($F_{12,4900} = 1.06$, $p = 0.392$), and the three-way interaction remains non-significant ($F_{48,4900} = 0.91$, $p = 0.652$). These results demonstrate that, although the macroscopic structural inference of sample profiles is sensitive to distinct interaction scales, the overall resolution power of the latent model remains strictly constrained by the primary generative parameters.

Consistent with the evaluation of the taxonomic compositions, the variance homogeneity of the sample-level prevalence reconstruction scores (Γ) is formally assessed. Levene's test is applied model to determine if the inference uncertainty is subject to heteroskedasticity depending on the generative parameters. The results (Table 5.4) indicate significant heteroskedasticity across all tested experimental factors. The dispersion parameter (ϕ) and the cohort size (N_g) exert an impact on the variance of the correlation scores, as evidenced by infinitesimal adjusted p-values. The biological effect amplitude (A) also significantly modulates this dispersion, albeit to a slightly lesser extent. These findings statistically corroborate the visual observations (Figure 5.8): structural overdispersion and sampling effort govern not only the mean inference fidelity but also the structural stability of the estimates across the simulations.

5.2.3 Taxon observation probability reconstruction (matrix P): RMSE evaluation

The global reconstruction fidelity is quantified by comparing the simulated generative probability matrices (P_{sim}) to the corresponding estimations inferred by the LDA model (P_{LDA}). This evaluation utilizes the RMSE across the full experimental design (Figure 5.9).

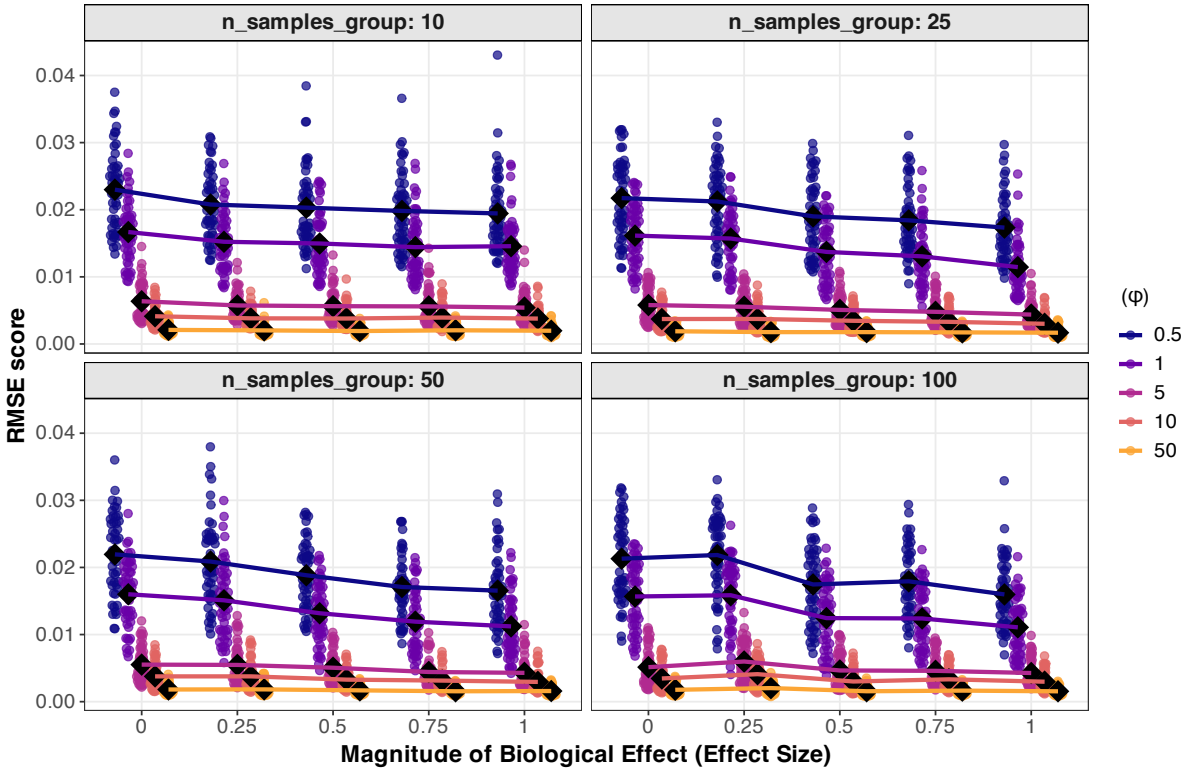


Figure 5.9: **Evolution of the taxon observation probability reconstruction errors (matrix P) across the experimental design.** The graph illustrates the distribution of the RMSE computed between the simulated ground truth matrices (P_{sim}) and the inferred matrices (P_{LDA}). A score approaching 0 indicates optimal structural reconstruction. Each point represents an independent simulation. The data are stratified into facets according to cohort size (N_g, ranging from 10 to 100 samples per group). The x-axis represents the magnitude of the imposed biological effect (A). Colors indicate the dispersion parameter ϕ (ranging from blue for high overdispersion to orange for low overdispersion). Black diamonds denote the mean error for each condition. Trend lines are colored according to ϕ.

In contrast to the previously employed similarity metrics, the RMSE scale is inversely proportional to reconstruction fidelity: a value approaching zero indicates better structural precision. The observed error scores span a restricted interval, ranging from approximately 0.04 (highest observed error) to 0.0007 (optimal precision). Consistent with the evaluation of the latent matrices (β and Γ), visual analysis displays that stochastic overdispersion (ϕ) remains the predominant factor driving inference degradation.

However, the variance distribution of this global error exhibits distinct structural differences when compared to the independent latent parameter estimations. Primarily,

the variance amplification effect induced by cohort size (N_g) under highly overdispersed conditions appears visually attenuated. Furthermore, whereas the latent similarity scores demonstrated a systematic global improvement as sampling increased, the global error does not display a corresponding systematic reduction pattern of its upper bound. The maximal errors generated under extreme overdispersion ($\phi = 0.5$) plateau at approximately 0.02, independent of the cohort size.

Conversely, the optimal precision threshold for minimal overdispersion conditions (the lower bound of the error distribution) is achieved efficiently within smallest cohorts tested ($N_g = 10$), and exhibits increased structural stability as sample sizes expand. Finally, the mean trend lines display increased slope variability within larger cohorts, indicating a potentially complex interaction effect of the biological amplitude (A) on the mitigation of the global structural error.

The preceding visual observations regarding the global reconstruction of the observation probabilities (P) are formally quantified through an analysis of deviance (Table 5.5). Given the continuous, strictly positive, and right-skewed nature of the error metric, this sensitivity analysis is derived from a GLM utilizing a Gamma error distribution and a log link function. (A full version of the table is provided in Appendix A.8.)

Table 5.5: **Sensitivity analysis of the probability reconstruction errors (matrix P) – Analysis of deviance.** The table presents the F-statistics and associated empirical p-values for the main effects and interactions evaluated across the experimental design ($N = 5000$ simulations). Degrees of freedom for the numerator and denominator (Residuals: 4900) are indicated.

Factor / Interaction	**F-value**	**p-value**
Dispersion Parameter (ϕ)	5612.09	< 0.001
Biological Effect Amplitude (A)	53.01	< 0.001
Cohort Size (N_g)	34.79	< 0.001
$A \times N_g$	4.10	< 0.001
$\phi \times A$	1.04	0.408
$\phi \times N_g$	0.41	0.959
$\phi \times N_g \times A$	0.16	1.000

The dispersion parameter (ϕ) consistently remains the dominant factor governing the variance of the reconstruction error ($F_{4,4900} = 5612.09$, $p < 0.001$). However, in contrast to the specific latent evaluations (β and Γ), a distinct shift in the relative impact of the secondary effects is observed. The biological effect amplitude A ($F_{4,4900} = 53.01$, $p < 0.001$) overtakes the cohort size N_g ($F_{3,4900} = 34.79$, $p < 0.001$). This shift demonstrates that the restoration of the global matrix structure relies more heavily on the initial strength of the cohort covariate than on the raw number of samples available to resolve the inference.

Examination of the interaction terms confirms the distinct structural dynamics of this global metric. Unlike the isolated latent parameter reconstructions, all interactions involving structural overdispersion ($\phi \times N_g$ and $\phi \times A$) entirely lose their statistical significance ($F_{12,4900} = 0.41$, $p = 0.959$ and $F_{16,4900} = 1.04$, $p = 0.408$, respectively). This statistical independence formally translates the previous visual observation: neither

increased sample sizes nor pronounced biological signals effectively mitigate the substantial global error generated under severe stochastic noise conditions.

Conversely, the joint relationship between the biological effect and the cohort size ($A \times N_g$) emerges as the sole significant two-way interaction ($F_{12,4900} = 4.10$, $p < 0.001$). Finally, the recurring absence of a significant three-way interaction ($F_{48,4900} = 0.16$, $p = 1.000$) mathematically indicates that the global reconstruction fidelity is not subjected to complex higher-order parameter interdependencies.

Beyond the evaluation of mean reconstruction scores, the dispersion of the fidelity estimates is formally assessed to determine if the inference uncertainty is subject to heteroskedasticity. Levene's test is conducted to quantify the influence of generative parameters on the variance of the similarity scores. The results (Table 5.6) reveal a structural divergence compared to the isolated latent matrices. While the dispersion parameter (ϕ) and the biological effect amplitude (A) exert a significant impact on the variance of the global error (with the influence of ϕ exceeding standard computational precision limits, $p < 2.2 \times 10^{-16}$), the cohort size (N_g) does not induce significant heteroskedasticity ($p = 0.181$). This statistical finding confirms the previous visual observation (Figure 5.9): although sample size influences the mean inference error, it does not significantly modulate the dispersion boundaries of that global error across independent simulations.

Table 5.6: **Impact of simulation parameters on the variance of probability reconstruction errors (P).** Results of Levene's tests evaluating the homogeneity of variance across experimental generative factors. Empirical p-values are adjusted using the FDR method to account for multiple comparisons. p-values computed as exactly zero by the software are reported as being below the machine epsilon ($< 2.2 \times 10^{-16}$).

Factor	**Raw p-value**	**Adjusted p-value (FDR)**
Biological Effect Size (A)	7.13×10^{-12}	1.07×10^{-11}
Size Parameter (ϕ)	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
Cohort Size (N_g)	0.181	0.181

5.2.4 Summary of Reconstruction Performance Across Latent Matrices

In conclusion, the systematic evaluation of inference fidelity demonstrates that the latent model possesses a reliable reconstruction capacity, strictly conditioned by the balance of the generative parameters. Across all evaluated metrics, the size parameter (ϕ) consistently constitutes the predominant limiting factor, governing both the variance and the overall precision of the estimations.

The analyses nevertheless reveal distinct parametric sensitivities depending on the reconstructed matrix. The precise structural resolution of the taxonomic compositions (β) depends on an interaction between stochastic overdispersion and sampling effort ($\phi \times N_g$), where expanded cohorts ($N_g = 100$) are necessary to stabilize the structural estimates. The inference of sample-level prevalences (Γ) shares this exact interaction dynamic ($\phi \times N_g$), yet the optimal fidelity plateau is achieved with more restricted cohort sizes ($N_g = 50$). Furthermore, a distinct shift in the relative impact of the main effects is

definitively confirmed during the evaluation of the global probability error (P). For this macroscopic metric, the biological effect amplitude (A) overtakes the cohort size (N_g) as the secondary influencing factor, while all interactions involving stochastic overdispersion lose their statistical significance.

Finally, the systematic absence of a significant three-way interaction across all deviance analyses is a favorable indicator of the structural stability of the inference framework. The joint influence of the three simulation parameters remains decomposable into interpretable main effects and pairwise interactions, suggesting that the degradation mechanisms operate independently rather than through complex entangled dynamics. This parametric separability ensures that the performance of the pipeline remains interpretable and predictable across the factorial design. Taken together, these results highlight two distinct optimization targets: cohort size primarily governs the precision of taxonomic signature recovery, while biological effect magnitude constitutes the principal determinant of successful cohort differentiation.

5.3 Type I error control and statistical power of PERMANOVA group testing

Beyond structural inference fidelity, this section assesses the preservation of the biological signal required for cohort differentiation. The analytical framework operates sequentially: the LDA algorithm first infers the latent space in a strictly unsupervised manner, without prior knowledge of sample affiliations. Subsequently, the predefined cohort covariates are explicitly provided to a PERMANOVA test. This formal statistical test assesses whether the variance explained by the group affiliations is significantly detectable within the inferred topic prevalences (matrix Γ). The reliability of this pipeline is quantified through a joint analysis of the false positive rate (Type I error) and the statistical power (true positive rate).

5.3.1 False positive rate under null conditions ($A = 0$): Type I error control

The evaluation of the False Positive Rate (FPR) is a fundamental requirement to validate the statistical reliability of the analytical pipeline. This specific analysis is strictly restricted to simulations lacking an underlying biological effect between the cohorts ($A = 0$). This step ensures that the combination of stochastic overdispersion (ϕ) and the latent inference process (LDA) does not generate artificial cohort separations in the absolute absence of a true group effect.

Visual inspection of the distribution (Figure 5.10) reveals the absence of a systematic pattern linking the generative parameters to an inflation of the Type I error. The measured FPR oscillates stochastically within a restricted boundary, ranging from 0% (observed under conditions such as $N_g = 25, \phi = 0.5$ and $N_g = 50, \phi = 10$) to a maximum of 8% (observed under conditions such as $N_g = 10, \phi = 1$ and $N_g = 50, \phi = 0.5$). Because these rates are computed across 50 independent replicates per factorial condition, the error percentages inherently fluctuate by discrete increments of 2%.

The observed error rates distribute consistently around the nominal 5% (p-value < 0.05)

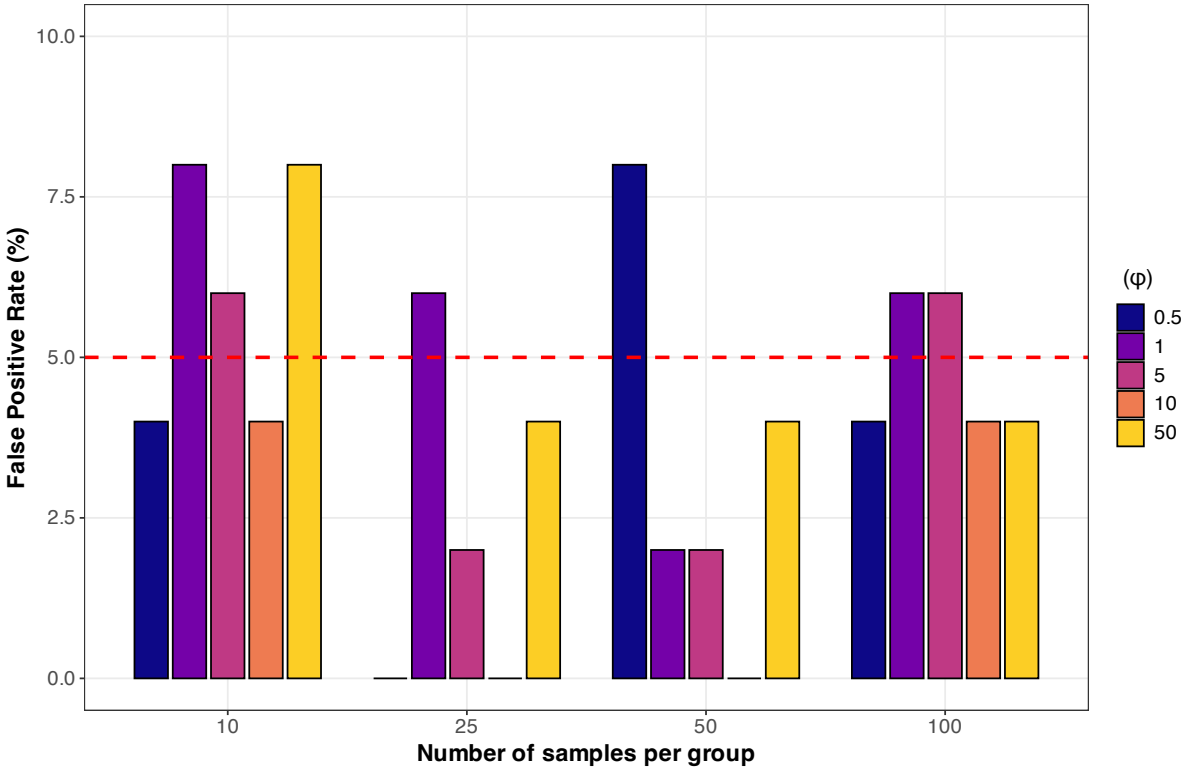


Figure 5.10: **Evaluation of the False Positive Rate (Type I error) across varying cohort sizes (N_g) and dispersion parameters (ϕ).** The bar chart illustrates the percentage of simulations, generated strictly without a true biological effect ($A = 0$), that yielded an erroneous statistically significant group effect ($p \leq 0.05$) during the PERMANOVA test applied to the inferred prevalence matrix (Γ_{LDA}). The red dashed line denotes the standard nominal tolerance threshold set at 5% (p-value < 0.05). The x-axis stratifies the results by cohort size (N_g), and the bar colors indicate the dispersion parameter ϕ (ranging from 0.5 in blue to 50 in yellow).

tolerance threshold (red dashed line). Specifically, 13 out of the 20 evaluated conditions (65%) remain strictly below the 5% mark. This stochastic distribution represents the expected statistical behavior of an unbiased hypothesis testing framework. Most notably, severe stochastic overdispersion (decrease in ϕ), which previously demonstrated a profound degradative impact on structural inference fidelity, does not induce spurious cohort separations.

The lack of systematic correlation between the dispersion parameter (ϕ), the cohort size (N_g), and the FPR supports the validity of the pipeline. It demonstrates that the unsupervised LDA algorithm does not translate extreme stochastic variance into artificial group structures. The reliability of the inference is consequently validated in the absence of an underlying signal, confirming a controlled Type I error rate.

To statistically evaluate the stability of the FPR and formally test the influence of the generative parameters on the Type I error rate, GLM (specifying a binomial family and a logit link function is fitted). Subsequently, an analysis of deviance utilizing a LR test is conducted to assess the global significance of the main effects (Table 5.7).

The analysis of deviance (Table 5.7 mathematically corroborates the stochastic distribution

Table 5.7: **Analysis of deviance evaluating the global impact of generative parameters on the Type I error rate.** The table presents the LR χ^2 statistics, degrees of freedom (Df), and associated p-values derived from the binomial GLM. This analysis assesses whether the dispersion parameter (ϕ) or cohort size (N_g) significantly alters the probability of generating a false positive under null biological effect conditions ($A = 0$).

Factor	**Df**	**LR χ^2**	**p-value**
Size Parameter (ϕ)	4	4.07	0.397
Cohort Size (N_g)	3	5.04	0.169
$N_g : \phi$	12	11.74	0.467

observed visually (Figure 5.10). The results confirm that neither structural overdispersion (ϕ) nor cohort size (N_g) exerts a statistically significant global impact on the probability of generating a false positive (all $p > 0.05$). This statistical independence definitively establishes that the LDA inference framework does not artificially inflate the PERMANOVA Type I error rate, even under extreme conditions of stochastic noise or restricted sampling. The detailed model coefficients, expressed as OR, are provided for exhaustive reference in Appendix A.9.

5.3.2 Statistical power of cohort effect detection across simulation parameters

Statistical power evaluates the sensitivity of the analytical framework, specifically its capacity to correctly detect a simulated biological effect when it is truly present ($A \neq 0$). Within this evaluation, a successful detection is formally defined by a PERMANOVA p-value strictly below the standard 0.05 significance threshold. Maximizing this metric is strictly equivalent to minimizing the risk of producing false negatives (Type II error, β), thereby ensuring that true underlying biological signals are successfully identified following the inference process.

The evaluation of the statistical power (Figure 5.11) reveals more structured detection dynamics than those observed during the FPR analysis (Figure 5.10). Globally, the capacity of the pipeline to correctly identify a biological difference increases concurrently with the three generative parameters: a higher biological signal amplitude (A), an expanded cohort size (N_g), and reduced stochastic overdispersion (increased ϕ).

Visual inspection (Figure 5.11) allows for the identification of several critical operational limits below which the inference fails to reach the 80% power target:

1. Severe stochastic overdispersion ($\phi = 0.5$) systematically suppresses detection, unless compensated by a highly pronounced biological signal ($A \geq 0.75$) **or** expanded cohort sizes ($N_g \geq 50$).
2. A biological effect of low magnitude ($A = 0.25$) remains largely undetectable, it necessitates minimal overdispersion ($\phi \geq 5$) and maximal sampling ($N_g = 100$) to achieve the significance threshold.
3. Restricted cohorts ($N_g = 10$) exhibit insufficient overall power, surpassing the 80% threshold exclusively under conditions of a clear biological signal ($A \geq 0.5$) coupled

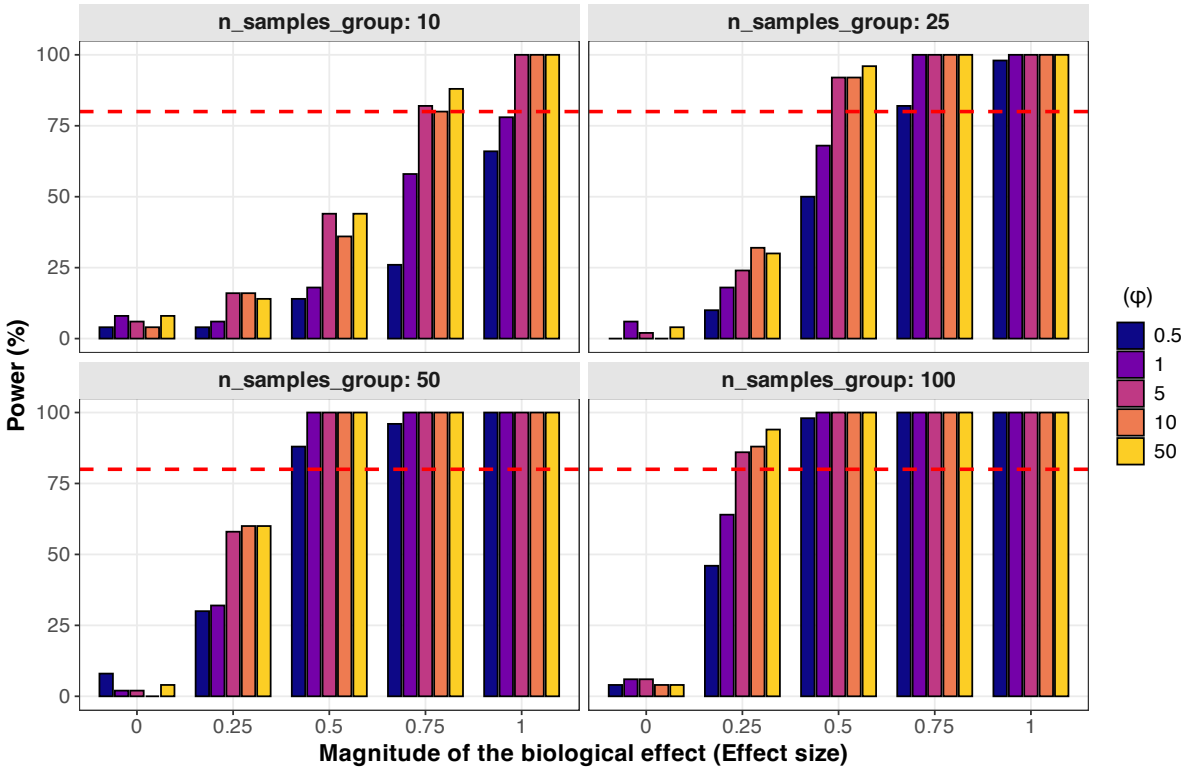


Figure 5.11: **Evaluation of the statistical power (1 - Type II error) across the experimental design.** The bar chart illustrates the percentage of simulations, that yielded a correct, statistically significant group effect ($p \leq 0.05$) via the PERMANOVA test applied to the inferred prevalence matrix (Γ_{LDA}). The red dashed line indicates the target statistical power threshold, classically established at 80%. The results are stratified into facets according to cohort size (N_g). The x-axis represents the magnitude of the imposed biological effect (A), and the bar colors denote the dispersion parameter ϕ (ranging from blue for high overdispersion to yellow for low overdispersion).

with controlled structural variance ($\phi \geq 5$).

Conversely, several factorial configurations allow a high-performance plateau (power $\geq 80\%$) to be consistently reached:

1. Cohort sizes of $N_g = 25$ mark a transition toward high detection rates, with statistical power consistently exceeding the 80% threshold provided that either the biological effect is moderate ($A \geq 0.25$) or stochastic overdispersion is well-controlled ($\phi \geq 5$).
2. Moderate to low overdispersion ($\phi \geq 5$) ensures high statistical power provided the biological effect amplitude is at least moderate ($A \geq 0.5$).
3. An effect amplitude of $A = 0.75$ allows for highly reliable detection across the vast majority of scenarios, rendering the results obtained at $A = 1$ nearly perfect. The achievement of these high performance levels empirically validates the upper boundary set for this generative parameter (exploring higher signal amplitudes would be superfluous for the evaluation of the framework's detection capacity).

To statistically quantify these compensatory dynamics and establish the relative importance of the generative factors, GLM with a binomial family and a logit link function is fitted.

Subsequently, an analysis of deviance utilizing a LR test is conducted to assess the global significance and effect size of each parameter and their interactions (Table 5.8).

Table 5.8: **Analysis of deviance evaluating the global impact of generative parameters on statistical power.** The table presents the LR χ^2 statistics, degrees of freedom (Df), and associated p-values derived from the binomial GLM. This analysis assesses the global significance of the biological effect amplitude (A), structural dispersion (ϕ), and cohort size (N_g).

Term	**LR χ^2**	**Df**	**p-value**
Effect Size (A)	1538.52	4	< 0.001
‘ Cohort Size (N_g)	1018.08	3	< 0.001
Size Parameter (ϕ)	267.31	4	< 0.001
$A : N_g$	72.37	9	< 0.001
$A : \phi$	39.94	12	< 0.001
$\phi : N_g$	14.34	12	0.279
$A : N_g : \phi$	15.93	36	0.998

The analysis of deviance corroborates the visual trends, identifying the biological signal amplitude (A) and the cohort size (N_g) as the primary determinants of statistical power. The magnitude of the OR (detailed in Appendix A.10) further confirms this shift in the relative magnitude of the generative factors: while a substantial increase in sampling effort ($N_g = 100$) multiplies the odds of detection by ~ 170 (OR: 169.48), the influence of the biological signal amplitude remains even more pronounced, with an OR of ~ 497 for the maximum effect ($A = 1.0$).

Notably, this analysis reveals a structural divergence from the previous reconstruction fidelity assessments: the dispersion parameter (ϕ) now constitutes the least influential factor governing detection power. While structural overdispersion remains a significant covariate, its relative impact on the probability of a successful detection is markedly inferior to that of the cohort size or the biological signal.

Furthermore, the significant interactions between the biological signal and the technical parameters ($A \times \phi$ and $A \times N_g$) validate the previously observed compensatory dynamics. These results demonstrate that the pipeline efficiency is primarily governed by the strength of the biological contrast, followed closely by the cohort sampling depth, with these parameters acting in synergy to mitigate the deleterious impact of structural overdispersion.

5.3.3 Summary of PERMANOVA Reliability: Type I Error and Statistical Power

In conclusion, the joint evaluation of Type I and Type II errors demonstrates the statistical reliability of the LDA inference coupled with a PERMANOVA test framework. The results indicate that the pipeline maintains a controlled FPR, oscillating around the nominal 5% threshold, even under extreme conditions of overdispersion or weak effect size. This control ensures that the latent inference process does not translate structural noise into spurious cohort separations. Conversely, the statistical power analysis highlights the parametric requirements for the accurate detection of true group-level differences.

Although the inference performance reaches operational limits when faced with maximal tested stochastic overdispersion ($\phi = 0.5$), restricted cohort sizes ($N_g = 10$), or minimal biological signals ($A = 0.25$), the framework exhibits compensatory dynamics. These results demonstrate the suitability of the LDA pipeline to preserve and extract the biological signal required for cohort differentiation. Specifically, the deleterious influence of stochastic overdispersion can be mitigated by either sufficient sampling effort (ideally $N_g \geq 50$) or, most effectively, by the presence of a pronounced biological contrast ($A \geq 0.75$).

6. Discussion

6.1 Summary of the Pipeline Performance

While the present simulations define experimental groups purely through mathematical shifts in latent topic prevalence, the ultimate objective of applied microbiota profiling is to reliably associate specific microbial structures with real-world phenotypes. As emphasized by Debelius et al. (2016), these biological cohorts can be defined by a diverse array of criteria, such as distinct dietary patterns, age gradients, or specific medication exposures. Large-scale population studies demonstrate that these clinical and environmental covariates act as deterministic drivers of microbiome composition, generating the macroscopic cohort signatures targeted by statistical inference (Zhernakova et al. 2016). However, successfully identifying these overarching cohort effects is notoriously difficult due to inherent data overdispersion, necessitating rigorous experimental design to secure adequate statistical power (Debelius et al. 2016).

Within this context, the primary objective of this study is to evaluate the capacity of an analytical pipeline, combining a LDA model and a PERMANOVA test, to infer these overarching differences between known *a priori* microbiota cohorts based on their latent topic prevalence profiles. Consequently, the main tasks consist of verifying the reliability of the LDA algorithm to reconstruct the simulated latent matrices (composition β, prevalence Γ, and taxon observation probabilities (P)), and subsequently assessing the statistical power of the PERMANOVA test to detect the simulated cohort effect from these inferred variables. To achieve this, a full factorial design is implemented, varying the biological effect amplitude (A, 5 levels), the overdispersion modeled by the size parameter (ϕ, 5 levels)[1], and the cohort size (N_g, 4 levels). This framework enables the identification of the experimental conditions influencing the fidelity of the latent inference and the statistical accuracy of the group comparison test.

The systematic analysis of the latent matrix reconstruction scores (β, Γ, and P) reveals common inference dynamics. The size parameter (ϕ) is consistently identified as the primary explanatory factor for the model deviance. Inference fidelity is substantially reduced by pronounced overdispersion (characterized by $\phi \leq 1$), resulting in decreased similarity scores and an increased RMSE. Although the other two evaluated main effects (A, N_g) demonstrate a statistically significant impact on reconstruction accuracy, their relative importance varies depending on the evaluated matrix. For the β and Γ matrices, cohort size (N_g) exerts a greater influence than biological effect magnitude (A), whereas a reversal of this relative importance is observed when evaluating the taxon observation probabilities (P). The predominant impact of overdispersion is further confirmed by the

[1] In the statistical literature, the coefficient controlling the variance term is alternately defined as a dispersion parameter (α, where $Var = \mu + \alpha\mu^2$, such that higher values increase variance) or a size parameter (ϕ, where $Var = \mu + \mu^2/\phi$, such that higher values decrease variance). The size parameterization is explicitly utilized here to directly align with its native implementation in the `rzinb` package used within the simulation codebase.

analysis of two-way interactions: the ϕ parameter is a central component of the most influential interactions for the latent compositions (namely $\phi \times N_g$ for the β matrix and $\phi \times A$ for the Γ matrix) (Tables 5.1, 5.3, 5.5). Furthermore, the absence of any significant three-way interaction indicates that main effects and two-way interactions are sufficient to capture the variance in algorithmic performance. Beyond average performance trends, the stability of the unsupervised inference is assessed by analyzing the heteroskedasticity of the reconstruction scores. Median-centered Levene's tests demonstrate that the dispersion of the metrics is significantly influenced by the three simulation parameters. Consistent with the sensitivity analysis, the overdispersion parameter (ϕ) acts as the primary driver of inference instability across all evaluated matrices. For the latent composition (β) and prevalence (Γ) matrices, score variance is additionally affected by cohort size (N_g) and biological effect amplitude (A). Conversely, the dispersion of the RMSE for taxon observation probabilities (P) is solely driven by overdispersion and biological effect magnitude, as cohort size does not yield a statistically significant impact on this specific variance.

Regarding the reliability of the group comparison, the analysis highlights the stability of the analytical pipeline in the absence of an underlying biological signal ($A = 0$). A primary concern when applying unsupervised algorithms to compositional data is the potential generation of spurious clustering driven entirely by sequencing noise and structural sparsity (Gloor et al. 2017). However, the stability of the Type I error rate indicates that the pipeline does not generate illusory cohort differences, even when confronted with extreme overdispersion. The FPR fluctuates stochastically around the 5% nominal threshold, ranging between 0% and 8% (Figure 5.10). This robust control of the FPR is particularly noteworthy, as recent large-scale benchmarking studies have demonstrated that traditional differential abundance methods frequently fail to control false discovery rates under the null hypothesis, often generating an alarming excess of spurious discoveries (Hawinkel et al. 2019; Thorsen et al. 2016).

Conversely, the analysis of statistical power reveals operational thresholds governing the capacity to detect the cohort effect (Figure 5.11). Statistical power is compromised ($< 80\%$) under conditions of restricted sampling ($N_g = 10$) combined with a weak biological signal ($A = 0.25$) or pronounced overdispersion ($\phi \leq 1$). Below these thresholds, underlying cohorts signatures are not sufficiently resolved within the inferred latent space to yield a statistically significant group difference. Alternatively, an operational plateau (power $\geq 80\%$) is reached under more favorable experimental parameters: attenuated overdispersion ($\phi \geq 5$), a marked biological effect ($A \geq 0.75$), or a sufficiently sized cohort ($N_g \geq 25$) consistently enhances the detection of the cohorts signatures. These empirical tipping points are supported by significant ORs derived from the GLM (Table A.10). Furthermore, under highly discriminative configurations combining optimal parameter values (e.g.,, strong biological signal and minimal overdispersion), the cohort profiles are distinctly separated within the inferred latent space. This maximal inter-group distance ensures the systematic detection of the underlying biological effect, reliably yielding 100% statistical power.

6.2 Combining LDA dimensionality reduction and PERMANOVA testing for microbiome cohort comparison

The application of distance-based statistical tests directly to raw microbiome count matrices presents significant mathematical challenges. Because sequencing data are inherently compositional (Gloor et al. 2017), highly dimensional ($p >> N$), profoundly sparse and subject to severe structural overdispersion (ϕ), traditional comparative analyses often lack statistical power and fail to control false discovery rates (Hawinkel et al. 2019). When the PERMANOVA is applied directly to the high-dimensional taxon space, the true biological effect (A) is frequently masked by stochastic noise and individual heterogeneity. The "curse of dimensionality" inherently distributes the statistical evaluation across microscopic fluctuations among hundreds of rare taxa, diluting the macroscopic biological variance and severely degrading the reliable detection of cohort signatures.

To circumvent these structural limitations, dimensionality reduction via LDA functions as a structural filter. Instead of treating taxa as independent variables, the LDA extracts deterministic co-occurrence patterns, grouping them into macroscopic latent sub-communities (topics). Crucially, unlike DMM models or global hard-clustering approaches that constrain each sample into a single discrete cluster (Costea et al. 2018; Ravel et al. 2011), LDA is a mixed-membership model (Blei et al. 2003; Sankaran and S. Holmes 2019). It models samples as exhibiting fractional memberships across multiple topics, an approach that fundamentally aligns with the ecological reality of the microbiota, where microbes operate as continuous interacting networks rather than mutually exclusive isolated entities (Pappalardo et al. 2024; Sankaran and S. Holmes 2019; Symul et al. 2023).

Recent applications confirm the relevance of this architecture. Pappalardo et al. (2024) demonstrated that LDA-based models accurately reconstruct complex taxonomic structures across diverse microbial niches. Furthermore, Symul et al. (2023) successfully applied mixed-membership topic models to identify latent sub-communities in the vaginal microbiota, revealing longitudinal transitions and disease-associated states that traditional global clustering approaches failed to capture.

By tracking multiscale topic alignments, as formalized by Fukuyama et al. (2021) via the `alto` framework, the latent space provides a mathematically stable representation of the underlying biological structures. Crucially, this architecture maintains high structural fidelity when confronted with background data heterogeneity—a dynamic directly analogous to the severe overdispersion (ϕ) evaluated in the present simulations. Fukuyama et al. (2021) demonstrated that evaluating topic coherence across varying resolutions facilitates the distinction of true latent sub-communities from unmodeled sample-to-sample variation. Furthermore, Huo et al. (2025) recently highlighted the capacity of LDA, coupled with PERMANOVA, to resolve inter-sample heterogeneity in pathological ecosystems. The model successfully decomposes overlapping bacterial communities to reveal intermediate disease states that remain undetected by standard linear analyses.

Consequently, applying the *a posteriori* PERMANOVA to the LDA-derived prevalence matrix (Γ), rather than the raw taxonomic counts, structurally enhances the statistical comparison. This synergy leverages the capacity of the LDA to isolate true co-occurrence patterns from structural overdispersion. The dimensionality reduction compresses the

feature space from hundreds of dispersed taxa (F) (where individual fluctuations are typically uninformative (Sankaran and S. Holmes 2019)) down to a restricted number of mathematically stable latent topics (K). By evaluating the variance of these stabilized macroscopic communities, the PERMANOVA amplifies the signal-to-noise ratio. This specific strategy of performing downstream statistical evaluations on the LDA latent space rather than raw counts is mathematically demonstrated to improve discriminatory performance by filtering stochastic dispersion (Blei et al. 2003), a principle recently validated in microbial ecology for testing community associations by Pappalardo et al. (2024). This methodological pairing explains the high statistical power observed in the present simulations when a sufficiently distinct biological effect is present, confirming the pipeline as a reliable tool for discerning complex ecological shifts in overdispersed clinical or experimental datasets (Huo et al. 2025).

6.3 Influence of the Experimental Parameters on Pipeline Performance

6.3.1 Structural Impact of the Invariant Parameters

The sensitivity analysis evaluates the variable parameters of the experimental design (A, N_g, ϕ). However, the performance of the LDA algorithm is also constrained by the invariant parameters of the generative architecture, which dictate the dimensionality of the inference problem (see Section 4.2.1).

The *a priori* definition of the number of latent topics ($K = 5$) constitutes a restricted simulation framework. In empirical studies, the optimal number of topics is unknown, and its estimation (model selection) constitutes a primary source of analytical uncertainty (Pappalardo et al. 2024). By providing the LDA algorithm with the exact generative value of K, the present experimental design prevents the structural forced merging or splitting of latent profiles inherently caused by dimensionality misspecification, a well-documented phenomenon in multiscale topic analysis (Fukuyama et al. 2021). Although severe overdispersion may still induce the confusion of these profiles, this initial alignment between the generative ground truth and the latent search space contributes to the stability of the Type I error rate. Consequently, the pipeline evaluates accurately bounded biological constructs rather than artificial partitions driven by dimensional constraints.

Secondly, taxonomic richness (the number of taxa, F) is maintained at 100. From a algorithmic perspective, expanding this bacterial vocabulary could theoretically provide additional data points to refine the reconstruction of the composition matrix (β). Nevertheless, microbiome ecosystems are inherently characterized by a heavy-tailed rank-abundance distribution (France et al. 2020), where the vast majority of species exist at extremely low abundances (a phenomenon termed the "rare biosphere" (I. Holmes et al. 2012)). Due to these distributional constraints, increasing F primarily generates a larger pool of extremely rare taxa with infinitesimal emission probabilities. In the context of probabilistic topic modeling, such sparsity creates significant mathematical constraints (Blei et al. 2003; I. Holmes et al. 2012). Excessive taxonomic complexity structurally degrades the reconstructive capacity of the model, as the macroscopic latent signal becomes obscured by a multitude of micro-variations that are generally uninformative (Hawinkel et al. 2019; Sankaran and S. Holmes 2019). To mitigate this, standard text-mining preprocessing rou-

tinely filters out extremely rare terms to stabilize inference (Grün and Hornik 2011). This principle translates directly to microbial ecology, where recent LDA applications explicitly advocate for the stringent filtering of rare OTUs or ASVs prior to dimensionality reduction. Removing this extreme sparsity prevents sequencing artifacts from driving the discovery of spurious sub-communities (Huo et al. 2025; Pappalardo et al. 2024). Consequently, restricting F in the present simulations facilitates the evaluation of mathematically stable ecological structures rather than algorithmic noise.

Finally, the configuration of the composition parameter ($\eta = 5$) plays a fundamental role in the topology of the generated data. This parameter constrains the Beta distribution governing the generative composition matrix ($\beta \sim Dir(\alpha_c)$, where $\alpha_c \sim Beta(1, \eta)$), which ultimately shapes the simulated count matrix. Specifically, setting $\eta = 5$ defines a distribution characterized by a low mean ($\mu \approx 0.167$) and low variance ($\sigma^2 \approx 0.02$ (Forbes et al. 2010a)). This yields small, tightly clustered α_c priors that force the Dirichlet distribution toward the edges and corners of the simplex, thereby generating the structural sparsity characteristic of biological sub-communities (Blei et al. 2003; Fukuyama et al. 2021). The sorting step further organizes the topics along a gradient of taxonomic concentration: Topic 1, assigned the smallest α_c values, is dominated by a single taxon, while Topic K, assigned the largest values, exhibits a comparatively more even distribution. Increasing this η parameter would push the probability mass onto an even more restricted number of ultra-dominant species, reducing the remaining taxa to negligible residual probabilities (Wallach et al. 2009). Although such extreme sparsity accurately models certain low-diversity ecosystems (such as the vaginal microbiota (France et al. 2020; Symul et al. 2023)), it structurally increases the difficulty of latent resolution. Because LDA fundamentally relies on identifying co-occurrence patterns (Sankaran and S. Holmes 2019; Pappalardo et al. 2024), reducing minority taxa to infinitesimal probabilities deprives the inference of crucial associative links. Under VEM optimization (Blei et al. 2003), the absolute frequency of a single dominant taxon systematically overwhelms the subtle correlational signals originating from the rare taxa. The mathematical distinction between latent profiles is thus forced to rely on negligible variations in the distribution tail, lacking the statistical power to drive accurate inference. Conversely, while decreasing the η parameter would artificially facilitate algorithmic reconstruction by generating uniform communities with visible structural links, such a configuration was discarded as it fails to reflect the fundamentally overdispersed and skewed nature of genuine microbiota ecosystems (I. Holmes et al. 2012).

6.3.2 Effect of Each Simulation Parameter on Inference Fidelity and Group Effect Testing

6.3.2.1 Overdispersion (ϕ) as the Primary Driver of Inference Degradation

Evaluated across five discrete levels, the size parameter (ϕ) intervenes exclusively during the final stage of the generative process (*cf.* Section 4.1.2): the transition from the ideal taxon observation probability matrix (P) to the final count matrix via the NB distribution. This specific mathematical distribution constitutes the standard framework for modeling the inherent structural overdispersion characterizing microbiota sequencing counts (Hawinkel et al. 2019; McMurdie and S. Holmes 2014). The resulting altered count matrix is subsequently provided to the LDA algorithm as a Document-Term Matrix (DTM).

The severe degradation of inference fidelity in the presence of pronounced overdispersion ($\phi \leq 1$) is directly explained by the mathematical nature of this perturbation. Specifically, decreasing the ϕ parameter structurally inflates the variance of the generated counts relative to their mean. This inflation exacerbates stochastic abundance fluctuations, accurately simulating both intrinsic biological heterogeneity (such as inter-individual variability) and severe technical biases (such as PCR amplification noise) (Debelius et al. 2016; Gloor et al. 2017). Because the LDA architecture fundamentally requires recurrent taxonomic co-occurrence patterns to infer latent profiles, extreme overdispersion obscures true biological associations beneath random fluctuations, including artificial abundance peaks or fortuitous absences (I. Holmes et al. 2012). Consequently, the resolution of the underlying ecological signal is severely compromised, forcing the optimization process to model technical noise rather than true biological variance (Hawinkel et al. 2019). This limitation in separating signal from noise explains the collapse of the matrix reconstruction scores (Tables 5.1, 5.3, 5.5), resulting in the structural merging of distinct latent profiles (as illustrated by the similarity matrices for $\phi = 0.5$ and $\phi = 50$, Figure 5.3).

Crucially, while previous methodological studies demonstrate that unmitigated data variability severely diminishes the accuracy of standard differential abundance methods in microbial ecology (Hawinkel et al. 2019), the comparative analysis of the GLMs reveals a functional decoupling between inference fidelity and statistical detectability within the proposed pipeline. Although structural overdispersion (ϕ) acts as the primary driver of latent reconstruction error, its impact on the statistical power of the subsequent PERMANOVA (Table 5.8) is superseded by the amplitude of the biological effect (A) and the cohort size (N_g). Aligning with the principles established by Debelius et al. (2016), this reversal indicates that a sufficiently distinct biological signal systematically maintains significant macroscopic inter-group distances within the inferred prevalence matrix (Γ), even when the microscopic taxonomic reconstruction is partially degraded by sequencing noise.

6.3.2.2 Cohort Size (N_g): Inference Improvement through Sampling Expansion

Evaluated across four discrete levels, consistent with the law of large numbers, an increase in cohort size (N_g) improves overall pipeline performance (Blei et al. 2003; Kelly et al. 2015). This trend is uniformly observed across both the optimization of the latent matrix reconstruction scores (Tables 5.1, 5.3, 5.5) and the increased capacity for statistical group effect detection (Table 5.8).

This improvement primarily derives from the co-occurrence identification mechanics of the LDA architecture (Blei et al. 2003). The reliable extraction of latent patterns requires a critical mass of cross-observations. Because each sample provides a defined vector of observed taxa, increasing N_g multiplies the absolute number of available data points. This expanded data volume provides the VEM optimization process with sufficient empirical evidence to reach mathematical convergence and stabilize the estimation of corpus-level parameters (Blei et al. 2003). Consequently, the inference generates more precise composition and prevalence matrices while mitigating the impact of individual stochastic fluctuations. This observation directly aligns with established methodological guidelines, which identify adequate sample size as a primary requirement to secure reliable inference (Debelius et al. 2016), and effectively offsets the high-dimensional data sparsity

characteristic of microbiota sequencing (I. Holmes et al. 2012).

Furthermore, beyond optimizing the unsupervised inference, cohort size functions as a critical mathematical buffer during the subsequent PERMANOVA evaluation. While this downstream test evaluates the variance explained by the group covariate, restricted sample sizes inherently lack the residual degrees of freedom required to reliably distinguish overarching cohort signatures from background intra-group variance (Anderson 2001). By expanding the cohort size, the test gains the necessary statistical power to overcome structural overdispersion, ensuring that macroscopic inter-group distances within the inferred latent space yield a statistically significant association (Kelly et al. 2015).

This dynamic contrasts with the structural consequences of increasing taxonomic richness (F). Expanding the feature space requires the partition of a finite probability space (the unit simplex modeled by the Dirichlet distribution) into infinitesimal fractions. Because sequencing depth is fundamentally constrained (Gloor et al. 2017), this mathematical fragmentation inherently generates a vast majority of zero counts, severely exacerbating data sparsity (I. Holmes et al. 2012). Conversely, increasing the cohort size (N_g) does not alter the internal probabilistic distribution of individual samples. Within the theoretical framework of LDA, expanding the cohort provides additional exchangeable, or conditionally independent, observations of the underlying latent space rather than merely replicating existing data (Blei et al. 2003). The estimation process therefore benefits from an absolute increase in available empirical information without undergoing further probability fragmentation, thereby avoiding the loss of reconstruction fidelity associated with high-dimensional sparsity.

6.3.2.3 Biological Effect Amplitude (A): Signal-to-Noise Ratio and Cohort Detectability

Similarly to the sampling expansion, an increase in the magnitude of the biological effect (evaluated across five discrete levels of A) generates an overall improvement in inference fidelity. This positive dynamic is observed both in the optimization of the latent matrix reconstruction scores (Tables 5.1, 5.3, 5.5) and in the maximization of statistical power during group effect detection (Table 5.8).

The analysis of the control scenarios ($A = 0$) isolates the baseline behavior of the LDA algorithm in the absence of a conditioning covariate. Under these null conditions, the group-specific effect weights are nullified. The prevalence of the latent topics is therefore drawn from an identical Dirichlet prior (α_p) across all samples, driven strictly by the intercept of the weight matrix (W). Consequently, no structural difference exists between the cohorts. The introduction of a strictly positive A parameter anchors the latent topic prevalence to specific group metadata (*cf.* Section 4.1.1.2). Mathematically, scaling these effect weights polarizes the probability distributions, increasing the inter-group distance within the generative prevalence matrix (Γ).

Because a pronounced biological effect magnitude (A) generates deeply segregated taxonomic profiles, it strengthens the recurrent co-occurrence patterns within the simulated count matrix. This intensified structural signal systematically facilitates the inference of the latent space by the unsupervised model (Sankaran and S. Holmes 2019). By providing the VEM optimization process with sharply defined community boundaries, algorithmic convergence is enhanced, significantly improving the mathematical resolution

of both the composition (β) and prevalence (Γ) matrices (Blei et al. 2003; Mimno and McCallum 2012). This mathematical dynamic explains the systematic increase in latent reconstruction fidelity observed in contrasted simulated cohorts (Tables 5.1, 5.3, 5.5).

Furthermore, the impact of this amplitude parameter (A) is more important during the *a posteriori* statistical group comparison. The analysis of deviance applied to the binomial GLM (Table 5.8)) identifies the A parameter as the primary predictor for successful cohort detection. This importance derives from the mathematical formulation of the PERMANOVA, which evaluates the ratio of inter-group variance (the divergence between cohorts) to intra-group variance (natural heterogeneity and overdispersion) (Anderson 2001). During data simulation, the A parameter modulates this inter-group variance. As demonstrated in benchmarking studies using simulated spike-ins, increasing the magnitude of a biological effect systematically improves statistical power (Thorsen et al. 2016). By elevating this amplitude, the injected biological signal generates a compositional divergence that exceeds the dispersion induced by the stochastic parameter (ϕ). Consequently, the A parameter amplifies the signal-to-noise ratio within the inferred latent structures. When the biological variance sufficiently offsets structural fluctuations, high statistical power is achieved (Debelius et al. 2016). This dynamic demonstrates that a pronounced biological effect size is conserved throughout the unsupervised dimensionality reduction performed by the LDA, enabling its reliable evaluation as explained variance by the *a posteriori* PERMANOVA (Kelly et al. 2015).

6.3.3 Compensatory Dynamics: Interactions Between Signal (A), Noise (ϕ), and Cohort Size (N_g)

The preceding analyses isolate the mechanical impact of each simulation parameter. However, in the context of microbiota sequencing data, these variables do not operate independently. The analysis of the two-way interactions reveals compensatory dynamics, demonstrating that specific adjustments in the experimental design preserve inference fidelity and statistical power despite unfavorable conditions.

6.3.3.1 Stochastic Noise Dilution Through Cohort Expansion ($\phi \times N_g$)

In the generative framework, the size parameter (ϕ) of the NB distribution models the overdispersion inherently found in microbiome data (Hawinkel et al. 2019). This variance inflation drives the generated read counts to stochastically deviate from their theoretical probabilities. When the cohort size is restricted (e.g.,, $N_g \leq 25$), the estimation lacks sufficient cross-sectional data. Consequently, the unsupervised inference lacks the statistical depth required to separate this stochastic dispersion from true biological co-occurrence patterns, mathematically modeling random abundance fluctuations as genuine latent sub-communities.

However, expanding the cohort volume effectively mitigates this vulnerability. By multiplying the number of observations, the random deviations statistically balance out at the population scale. This structural dilution of overdispersion prevents the optimization process from modeling spurious variations, ensuring that the VEM algorithm converges toward the true underlying taxonomic probabilities (Debelius et al. 2016). This dynamic explains why the $\phi : N_g$ interaction emerges as the most significant interaction governing the reconstruction fidelity of both the composition (β) and prevalence (Γ) latent matrices:

an expanded sampling volume provides the mathematical depth necessary to restore the structural co-occurrence patterns otherwise masked by severe dispersion.

6.3.3.2 Noise Compensation Using Signal Amplitude ($\phi \times A$)

In a dynamic conceptually related to cohort expansion, the interaction between overdispersion (ϕ) and biological effect magnitude (A) illustrates a secondary compensatory mechanism. Unlike the $\phi \times N_g$ interaction, which impacts the entire latent space, the analysis of deviance identifies $\phi : A$ as the least influential significant interaction, restricting its impact exclusively to the reconstruction of the prevalence matrix (Γ) (Table 5.3). While an increase in N_g mitigates dispersion through statistical dilution, an expansion of the A parameter counterbalances stochasticity by amplifying the structural signal-to-noise ratio.

When data topology is degraded by pronounced overdispersion ($\phi \leq 1$), latent prevalence patterns are partially obscured by stochastic abundance fluctuations. However, configuring a high amplitude parameter ($A \geq 0.75$) introduces a distinct mathematical polarization between the simulated groups. This divergence inherently widens the inter-group distance relative to the intra-group variance induced by the ϕ parameter. Consequently, even among dispersed individual counts, the macroscopic group signature remains sufficiently contrasted. This structural polarization explains why a pronounced biological effect partially preserves the fidelity of the inferred Γ matrix, enabling the algorithm to approximate the true topic distributions despite suboptimal technical conditions.

6.3.3.3 The Additive Synergy of Sampling and the Biological Signal ($N_g \times A$)

The interaction between cohort size (N_g) and biological effect amplitude (A) structurally combines the benefits identified previously: sampling expansion refines algorithmic precision through increased data volume, while the amplitude parameter mathematically polarizes the latent profiles between groups. Consequently, the combination of an expanded cohort and a pronounced biological signal yields the most favorable inference conditions within the study.

However, the analyses of deviance, reveals a counter-intuitive result: although it combines the two primary improvement factors, the $N_g \times A$ interaction never emerges as a dominant explanatory variable. Furthermore, this interaction remains strictly non-significant for the reconstruction of the latent matrices (β and Γ, Tables 5.1, 5.3), reaching statistical significance only for the expected probability matrix (P) (Table 5.5). This observation is mathematically justified by the statistical properties of the evaluated models. First, the isolated main effects of N_g and A are intrinsically strong, thereby absorbing the majority of the explainable variance. Their synergy is thus additive rather than multiplicative: although higher values of these parameters mutually reinforce inference precision, their combination does not generate a disproportionate statistical effect beyond the sum of their individual marginal contributions. Second, the relative importance of the interaction terms is structurally dominated by factors involving overdispersion ($\phi \times N_g$ and $\phi \times A$). Because the variance inflation induced by ϕ constitutes the primary driver of inference degradation, compensatory dynamics (where an expanded sample volume or a high biological contrast restores an inference previously collapsed by severe dispersion) generate substantially stronger statistical weightings than the simple addition of two already favorable conditions.

6.3.4 Synthesis of Inference Dynamics

The fidelity of the LDA model is strictly conditioned by the characteristics of the input data: overdispersion (ϕ), biological effect magnitude (A), and cohort size (N_g). The invariant parameters of the experimental design, such as taxonomic richness (F) and specific distribution shapes (η), also exert a continuous influence on the results, the precise evaluation of which will require further investigation.

Among the evaluated variables, overdispersion (ϕ) acts as the primary degradation factor during the reconstruction of the latent probability matrices (β, Γ, and P). The high sensitivity of the latent inference to this stochastic dispersion, which is inherent to genuine microbiota sequencing data, warrants particular attention. However, the results demonstrate that this loss of precision is statistically compensable. Increasing the cohort size (N_g) dilutes the variance through the multiplication of available data points. Simultaneously, the biological signal magnitude (A) emerges as the decisive parameter for the *a posteriori* PERMANOVA. The reliable identification of the cohort effect demonstrates that the analytical pipeline successfully isolates authentic biological signatures without them being systematically masked by severe structural dispersion. Ultimately, the combination of adequate cohort sizing and a distinct biological signal provides favorable inference conditions, establishing the foundation for practical recommendations regarding the application of this methodology.

6.4 Limitations of the Simulation Framework and Future Perspectives

A primary methodological limitation of the current study lies in the restriction of the generative parametric space. To strictly isolate and quantify the influence of overdispersion (ϕ), cohort size (N_g), and biological effect magnitude (A), several structural variables are deliberately maintained constant.

First, the taxonomic richness is restricted ($F = 100$) to exclude extremely rare taxa, mirroring the standard bioinformatic practice of filtering raw sequencing count matrices. Future frameworks could adapt this parameter to match specific analytical requirements or varying filtering thresholds (Hawinkel et al. 2019; Huo et al. 2025). Simultaneously, the hyperparameter controlling baseline sparsity (η) is fixed to establish a single common baseline across all experimental conditions. However, this invariant η parameter prevents the evaluation of the LDA limits when confronted with extreme sparsity, a condition where the lack of shared taxa degrades the structural co-occurrences required to infer latent topics (Blei et al. 2003; Pappalardo et al. 2024). A direct perspective to address this limitation could involve transitioning from fixed theoretical values to data-driven empirical priors. As initially described in the foundational LDA framework (Blei et al. 2003), empirical Bayes parameter estimation could be implemented to directly infer realistic η parameters from reference microbiota datasets, thereby anchoring the generative sparsity strictly to genuine biological distributions.

Beyond the parametric constraints governing data sparsity, the topological complexity of the simulated cohorts is further constrained to a design with three groups ($G = 3$) and a predefined number of five latent sub-communities ($K = 5$). This controlled dimensionality serves to prevent confounding the inference degradation caused by the size parameter (ϕ)

with structural misspecification errors. However, the true number of biological topics (K) is rarely known *a priori* (Grün and Hornik 2011; Pappalardo et al. 2024). Consequently, it remains necessary to evaluate the pipeline's behavior when the LDA model is supplied with an incorrect K value during simulation. As demonstrated by Fukuyama et al. (2021), such parametric misspecification mathematically induces the artificial fusion of distinct sub-communities or the over-fractionation of continuous biological signals. A direct perspective to address this uncertainty involves the integration of a formal model evaluation step prior to the *a posteriori* PERMANOVA. While classical selection procedures rely on metrics such as predictive perplexity computed via cross-validation to determine a single optimal latent dimensionality (Grün and Hornik 2011), recent methodological frameworks recommend a multi-scale analysis (Fukuyama et al. 2021). This procedure requires executing the generative model across a range of candidate K values and assessing diagnostic metrics (such as topic coherence and refinement) to systematically track how biological signals align across dimensions and ensure the fidelity of the inferred signatures.

The univariate structure of the current generative framework constitutes an additional constraint. At present, a single categorical covariate (the cohort affiliation) conditions the latent space, and this effect is exclusively restricted to the simulated topic prevalence matrix (G). Consequently, the internal taxonomic composition of the topics (β) remains strictly unconditioned. To better reflect the multifactorial nature of genuine microbiota, future iterations require the integration of multivariate experimental designs. Incorporating additional covariates (e.g., age or technical factors) would allow for confounder adjustment directly during the latent inference (Mimno and McCallum 2012; Roberts et al. 2019; Valle et al. 2021). Furthermore, extending this parameterization to directly modulate the composition matrix (β) would represent an analytical evolution. Allowing biological or environmental factors to structurally alter the taxonomic probabilities within the sub-communities could significantly refine the identification of condition-specific biomarkers.

Beyond the immediate extensions of the current framework, a broader developmental perspective concerns the nature of the data structure itself. The current methodology is strictly conceptualized for cross-sectional data, where samples are assumed to be independent and temporally exchangeable (Blei et al. 2003). Extending this statistical framework to encompass longitudinal study designs therefore constitutes a natural yet substantial analytical evolution. Implementing architectures such as Dynamic Topic Models (Blei and Lafferty 2006) permits the temporal tracking of microbial sub-communities and their interactions. While transitioning to temporal dynamics represents fundamental methodological extension beyond the immediate scope of this study, such models provide the analytical resolution required to evaluate microbiota trajectories and structural resilience over time.

7. Conclusion

By evaluating the LDA-PERMANOVA pipeline through a controlled *in silico* full factorial design, this thesis establishes the parametric boundaries governing its reliability for detecting cohort differences in overdispersed microbiota composition data. The *in silico* framework developed here provides a structured, ground-truth-anchored environment in which the statistical behavior of the pipeline is systematically characterized, independently of the confounding factors inherent to empirical datasets. The full factorial design enables the precise isolation of the individual contributions of ϕ, N_g, and A to the overall performance of the pipeline.

While Huo et al. (2025) recently demonstrated the capacity of LDA to resolve severe inter-sample heterogeneity in empirical microbiota datasets, to our knowledge, the formal simulation-based characterization of the joint LDA-PERMANOVA pipeline remains largely unaddressed in the literature. The present work addresses this methodological gap. The functional decoupling between latent inference fidelity and statistical detectability constitutes a key finding of this evaluation: the framework exhibits reliable cohort differentiation capacity even when the latent taxonomic reconstruction is partially degraded by overdispersion. This stability represents a fundamental property that standard taxon-level approaches frequently struggle to maintain (Hawinkel et al. 2019; Thorsen et al. 2016).

The parametric boundaries established here define a foundation for subsequent methodological refinements, including the integration of K misspecification, multivariate covariates, and longitudinal architectures, as outlined in the preceding discussion. Consequently, this thesis positions the LDA-PERMANOVA framework as a mathematically principled alternative for navigating the inherent complexity of microbiota composition datasets, providing an approach whose operational limits are now quantitatively mapped.

Appendices

A. Supplementary Material

A.1 Supplementary Material: Computational Environment and Package Dependencies

All computational analyses are conducted in R (version 4.4.1) using RStudio (version 2026.05.0+218, "Golden Wattle"). The following table details the versions of the packages utilized throughout this pipeline to ensure reproducibility of the statistical analyses and graphical outputs.

Table A.1: **List of R packages and versions used in the analysis.** The table details the versions of the packages used throughout this study.

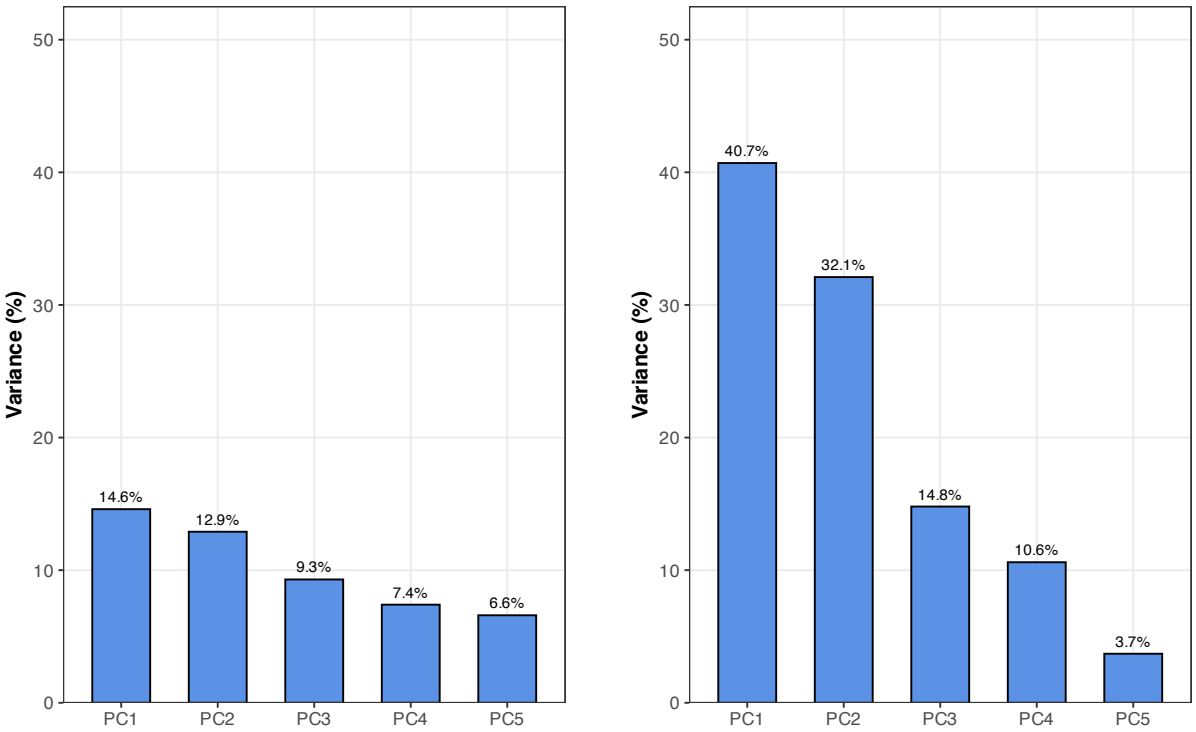
Package	Version
alto	0.1.0
broom	1.0.12
car	3.1-5
clue	0.3-67
DirichletReg	0.7-2
dplyr	1.2.0
forcats	1.0.1
Formula	1.2-5
ggplot2	4.0.2
ggbeeswarm	0.7.3
ggpubr	0.6.3
ggrepel	0.9.8
gt	1.3.0
httr	1.4.8
lubridate	1.9.5
magrittr	2.0.4
NLP	0.3-2
pander	0.6.6
patchwork	1.3.2
permute	0.9-10
pheatmap	1.0.13
philentropy	0.10.0
purrr	1.2.1
readr	2.2.0
rstatix	0.7.3
slam	0.1-55
stringr	1.6.0

Continued on next page

Table A.1 – continued

Package	**Version**
tibble	3.3.1
tidyr	1.3.2
tidyverse	2.0.0
tm	0.7-16
topicmodels	0.2-17
vegan	2.7-3
viridis	0.6.5
viridisLite	0.4.3
ZIM	1.1.0

A.2 Supplementary Results: Pipeline Validation



(a) High-Noise Scenario ($\phi = 0.5$)

(b) Low-Noise Scenario ($\phi = 50$)

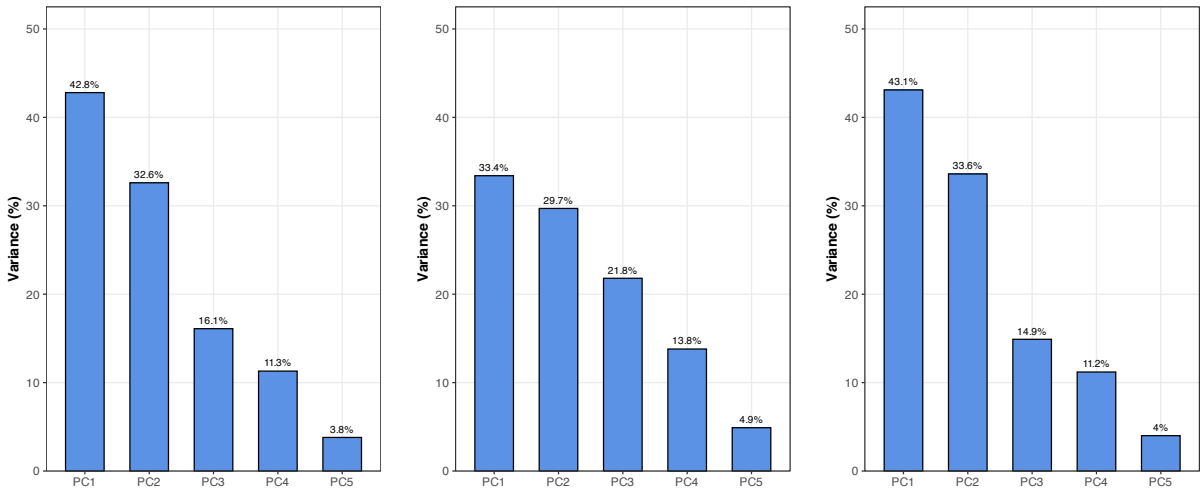
Figure A.1: **Variance explained by the first five principal components derived from the simulated relative abundance matrices.** The bar charts display the percentage of total variance (Y-axis) captured by each of the top five principal components (X-axis) following a Principal Coordinates Analysis (PCoA) based on Bray-Curtis dissimilarities. The left panel (A.1a) illustrates the variance distribution under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$). The right panel (A.1b) illustrates the variance distribution under minimal overdispersion (Low-Noise Scenario, $\phi = 50$). The biological effect amplitude ($A = 1$) and the cohort size ($N_g = 50$) are maintained constant in both scenarios. The blue bars represent the specific percentage of variance explained by each axis, visually quantifying the structural dilution of the biological signal caused by stochastic noise.

Table A.2: **PERMANOVA test results evaluating the cohort signal within the simulated relative abundance matrices.** Permutational Multivariate Analysis of Variance (PERMANOVA) was performed using Bray-Curtis dissimilarities and 999 free permutations to quantify the proportion of variance (R^2) explained by the predefined "Group" covariate prior to any latent inference. Results are presented for both the High-Noise ($\phi = 0.5$) and Low-Noise ($\phi = 50$) scenarios.

Scenario	**Pseudo-F**	R^2	p**-value**
High-Noise ($\phi = 0.5$)	2.906	0.038	0.001
Low-Noise ($\phi = 50$)	31.737	0.301	0.001

Table A.3: **Mantel test statistics evaluating the global structural preservation of sample-level topic prevalences (Γ).** The table presents the Mantel correlation coefficient (r) and the associated empirical p-value for both the High-Noise ($\phi = 0.5$) and Low-Noise ($\phi = 50$) scenarios. Tests were computed using Pearson's product-moment correlation over 999 free permutations, comparing the pairwise distance matrices derived from the simulated generative prevalences (G_{sim}) and the inferred proportions (Γ_{LDA}). The mean diagonal Spearman rank correlation scores are included for reference.

Scenario	Mean Spearman	Mantel Statistic (r)	p-value
$\phi = 0.5$	0.30	0.3019	0.001
$\phi = 50$	0.99	0.9802	0.001



(a) Ground Truth (P_{sim}) (b) (P_{LDA}, $\phi = 0.5$) (c) (P_{LDA}, $\phi = 50$)

Figure A.2: **Variance explained by the first five principal components derived from the generative and inferred taxon observation probability matrices (P).** The bar charts display the percentage of total variance (Y-axis) captured by each of the top five principal components (X-axis) following a Principal Coordinates Analysis (PCoA) based on Bray-Curtis dissimilarities. The left panel (A.2a) presents the variance distribution of the simulated ground truth (P_{sim}), which serves as the structural baseline. The middle panel (A.2b) illustrates the inferred matrix (P_{LDA}) under severe structural overdispersion (High-Noise Scenario, $\phi = 0.5$). The right panel (A.2c) illustrates the inferred matrix under minimal overdispersion (Low-Noise Scenario, $\phi = 50$). The biological effect amplitude ($A = 1$) and the cohort size ($N_g = 50$) are maintained constant across all represented conditions.

A.3 Supplementary Results: Factorial Design Sensitivity Analyses

Table A.4: **Quantitative reconstruction metrics for the taxon observation probability matrices (P).** The table presents the Root Mean Square Error (RMSE) computed element-wise between the simulated ground truth (P_{sim}) and inferred (P_{LDA}) probability matrices, alongside the Mantel correlation coefficient (r) and its associated empirical p-value. Mantel tests were computed using Pearson's product-moment correlation over 999 free permutations to evaluate global structural preservation across both the High-Noise ($\phi = 0.5$) and Low-Noise ($\phi = 50$) scenarios.

Scenario	**RMSE**	**Mantel Statistic (r)**	**p-value**
$\phi = 0.5$	0.021	0.6855	0.001
$\phi = 50$	0.003	0.9953	0.001

Table A.5: **PERMANOVA test results evaluating the preservation of the cohort signal within the sample-level topic prevalences (Γ).** Permutational Multivariate Analysis of Variance (PERMANOVA) was performed using Bray-Curtis dissimilarities and 999 free permutations to quantify the proportion of variance (R^2) explained by the predefined "Group" covariate. The generative ground truth (G_{sim}) serves as the structural baseline, strictly invariant to the dispersion parameter (ϕ). The inferred matrices (Γ_{LDA}) represent the reconstructed profiles under High-Noise ($\phi = 0.5$) and Low-Noise ($\phi = 50$) scenarios.

Matrix	**Size Parameter**	**Pseudo-F**	**R^2**	**p-value**
Generative Baseline (G_{sim})	Invariant	29.053	0.283	0.001
Inferred Prevalences (Γ_{LDA})	High-Noise ($\phi = 0.5$)	5.205	0.066	0.001
Inferred Prevalences (Γ_{LDA})	Low-Noise ($\phi = 50$)	29.222	0.284	0.001

Table A.6: **Sensitivity analysis of the taxonomic composition reconstruction scores (β) – Detailed analysis of deviance.** The table presents the full output of the Analysis Of Variance (ANOVA) model, including the Sum of Squares (*Sum Sq*), degrees of freedom (*Df*), F-statistics, and associated empirical p-values. The model evaluates the impact of the dispersion parameter (ϕ), cohort size (N_g), and biological effect amplitude (A) on the taxonomic reconstruction fidelity.

Factor / Interaction	**Sum Sq**	**Df**	**F-value**	**p-value**
ϕ	257.15	4	2975.32	< 0.001
N_g	32.35	3	499.06	< 0.001
A	7.11	4	82.24	< 0.001
$\phi \times N_g$	0.61	12	2.37	0.005
$\phi \times A$	0.48	16	1.39	0.134
$N_g \times A$	0.77	12	2.99	< 0.001
$\phi \times N_g \times A$	0.55	48	0.53	0.967
Residuals	105.87	4900	–	–

A.4 Supplementary Results: Statistical Modeling

Table A.7: **Sensitivity analysis of the sample-level prevalence reconstruction scores (Γ) – Detailed analysis of deviance.** The table presents the full output of the ANOVA model, including the Sum of Squares (*Sum Sq*), degrees of freedom (*Df*), F-statistics, and associated empirical p-values. The model evaluates the impact of the dispersion parameter (ϕ), cohort size (N_g), and biological effect amplitude (A) on the prevalence reconstruction fidelity (Spearman correlation).

Factor / Interaction	**Sum Sq**	**Df**	**F-value**	**p-value**
ϕ	570.38	4	4247.43	< 0.001
N_g	12.39	3	122.97	< 0.001
A	4.09	4	30.45	< 0.001
$\phi \times N_g$	5.13	12	12.72	< 0.001
$\phi \times A$	1.09	16	2.03	0.009
$N_g \times A$	0.43	12	1.06	0.392
$\phi \times N_g \times A$	1.46	48	0.91	0.652
Residuals	164.50	4900	–	–

Table A.8: **Sensitivity analysis of the probability reconstruction errors (P) – Detailed analysis of deviance.** The table presents the full output of the ANOVA model (derived from a GLM with a Gamma family and log link function), including the Sum of Squares (*Sum Sq*), degrees of freedom (*Df*), F-statistics, and associated empirical p-values. The model evaluates the impact of the dispersion parameter (ϕ), cohort size (N_g), and biological effect amplitude (A) on the global inference error (RMSE).

Factor / Interaction	**Sum Sq**	**Df**	**F-value**	**p-value**
ϕ	3554.81	4	5612.09	< 0.001
N_g	16.53	3	34.79	< 0.001
A	33.58	4	53.01	< 0.001
$\phi \times N_g$	0.79	12	0.41	0.959
$\phi \times A$	2.64	16	1.04	0.408
$N_g \times A$	7.80	12	4.10	< 0.001
$\phi \times N_g \times A$	1.19	48	0.16	1.000
Residuals	775.94	4900	–	–

Table A.9: **Generalized Linear Model (GLM) evaluating the impact of simulation parameters on the Type I error rate.** The table presents the Odds Ratios (OR) and their 95% Confidence Intervals (CI) relative to the baseline condition ($\phi = 0.5$, $N_g = 10$).

Parameter	Odds Ratio (OR)	95% CI	p-value
(Intercept)	0.062	[0.025, 0.133]	< 0.001
size (ϕ)			
$\phi = 1$	1.400	[0.553, 3.697]	0.481
$\phi = 5$	1.000	[0.360, 2.777]	> 0.999
$\phi = 10$	0.489	[0.128, 1.581]	0.249
$\phi = 50$	1.265	[0.487, 3.387]	0.629
Cohort Size (N_g)			
$N_g = 25$	0.384	[0.135, 0.963]	0.052
$N_g = 50$	0.517	[0.204, 1.214]	0.140
$N_g = 100$	0.789	[0.354, 1.722]	0.553

Table A.10: **Generalized Linear Model (GLM) evaluating the impact of simulation parameters on the statistical power of the model.** The table presents the Odds Ratios (OR) and their 95% Confidence Intervals (CI) relative to the baseline condition ($A = 0.25$, $\phi = 0.5$, $N_g = 10$).

Parameter	Odds Ratio (OR)	95% CI	p-value
(Intercept)	0.006	[0.004, 0.009]	< 0.001
Biological Effect (A)			
$A = 0.5$	19.596	[14.464, 26.879]	< 0.001
$A = 0.75$	102.662	[69.753, 154.046]	< 0.001
$A = 1.0$	496.818	[298.288, 856.910]	< 0.001
size (ϕ)			
$\phi = 1$	2.366	[1.722, 3.261]	< 0.001
$\phi = 5$	7.905	[5.568, 11.320]	< 0.001
$\phi = 10$	8.042	[5.661, 11.525]	< 0.001
$\phi = 50$	9.764	[6.822, 14.109]	< 0.001
Cohort Size (N_g)			
$N_g = 25$	10.615	[7.813, 14.576]	< 0.001
$N_g = 50$	44.192	[30.621, 64.819]	< 0.001
$N_g = 100$	169.484	[109.075, 269.186]	< 0.001

Bibliography

Airoldi, Edoardo M., David Blei, Elena A. Erosheva, and Stephen E. Fienberg, eds. (Nov. 6, 2014). *Handbook of Mixed Membership Models and Their Applications*. 0th ed. Chapman and Hall/CRC. ISBN: 978-0-429-09882-6. DOI: 10.1201/b17520. URL: https://www.taylorfrancis.com/books/9781466504097 (visited on 02/12/2026).

Aitchison, J. (Apr. 1, 1983). "Principal component analysis of compositional data". In: *Biometrika* 70.1, pp. 57–65. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.57. URL: https://doi.org/10.1093/biomet/70.1.57 (visited on 03/20/2026).

Anderson, Marti J. (2001). "A new method for non-parametric multivariate analysis of variance". In: *Austral Ecology* 26.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.2001.01070.pp.x, pp. 32–46. ISSN: 1442-9993. DOI: 10.1111/j.1442-9993.2001.01070.pp.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.2001.01070.pp.x (visited on 06/02/2026).

– (2017). "Permutational Multivariate Analysis of Variance (PERMANOVA)". In: *Wiley StatsRef: Statistics Reference Online*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118 John Wiley & Sons, Ltd, pp. 1–15. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat07841. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat07841 (visited on 03/20/2026).

Anderson, Marti J. and Daniel C. I. Walsh (2013). "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?" In: *Ecological Monographs* 83.4, pp. 557–574. DOI: 10.1890/12-2010.1. URL: https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-2010.1 (visited on 06/02/2026).

Arumugam, Manimozhiyan, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork (May 2011). "Enterotypes of the human gut microbiome". In: *Nature* 473.7346, pp. 174–180. ISSN: 1476-4687. DOI: 10.1038/nature09944. URL: https://www.nature.com/articles/nature09944 (visited on 12/06/2025).

Belkaid, Yasmine and Oliver J. Harrison (Apr. 18, 2017). "Homeostatic Immunity and the Microbiota". In: *Immunity* 46.4, pp. 562–576. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2017.04.008. URL: https://www.sciencedirect.com/science/article/pii/S1074761317301413 (visited on 05/22/2026).

Berche, Patrick (2007). *Une histoire des microbes*. In collab. with Raphaèle Dorniol. Montrouge, France ; Surrey, England: John Libbey Eurotext. 1 p. ISBN: 978-2-7420-0674-8.

Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter (June 30, 2020). "Microbiome definition re-visited: old concepts and new challenges". In: *Microbiome* 8.1, p. 103. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00875-0. URL: https://doi.org/10.1186/s40168-020-00875-0 (visited on 12/06/2025).

Blei, David and John D. Lafferty (2006). "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. the 23rd international conference. Pittsburgh, Pennsylvania: ACM Press, pp. 113–120. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143859. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143859 (visited on 06/07/2026).

Blei, David, Andrew Ng, and Michael Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research*, p. 30.

Canny, John (July 25, 2004). "GaP: a factor model for discrete data". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR04: The 27th ACM/SIGIR International Symposium on Information Retrieval 2004. Sheffield United Kingdom: ACM, pp. 122–129. ISBN: 978-1-58113-881-8. DOI: 10.1145/1008992.1009016. URL: https://dl.acm.org/doi/10.1145/1008992.1009016 (visited on 05/21/2026).

Costea, Paul, Falk Hildebrand, Manimozhiyan Arumugam, Fredrik Bäckhed, Martin J. Blaser, Frederic D. Bushman, Willem M. de Vos, S. Dusko Ehrlich, Claire M. Fraser, Masahira Hattori, Curtis Huttenhower, Ian B. Jeffery, Dan Knights, James D. Lewis, Ruth E. Ley, Howard Ochman, Paul W. O'Toole, Christopher Quince, David A. Relman, Fergus Shanahan, Shinichi Sunagawa, Jun Wang, George M. Weinstock, Gary D. Wu, Georg Zeller, Liping Zhao, Jeroen Raes, Rob Knight, and Peer Bork (Jan. 2018). "Enterotypes in the landscape of gut microbial community composition". In: *Nature Microbiology* 3.1, pp. 8–16. ISSN: 2058-5276. DOI: 10.1038/s41564-017-0072-8. URL: https://www.nature.com/articles/s41564-017-0072-8 (visited on 05/22/2026).

Costea, Paul, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, Marja Driessen, Rajna Hercog, Ferris-Elias Jung, Jens Roat Kultima, Matthew R. Hayward, Luis Pedro Coelho, Emma Allen-Vercoe, Laurie Bertrand, Michael Blaut, Jillian R. M. Brown, Thomas Carton, Stéphanie Cools-Portier, Michelle Daigneault, Muriel Derrien, Anne Druesne, Willem M. de Vos, B. Brett Finlay, Harry J. Flint, Francisco Guarner, Masahira Hattori, Hans Heilig, Ruth Ann Luna, Johan van Hylckama Vlieg, Jana Junick, Ingeborg Klymiuk, Philippe Langella, Emmanuelle Le Chatelier, Volker Mai, Chaysavanh Manichanh, Jennifer C. Martin, Clémentine Mery, Hidetoshi Morita, Paul W. O'Toole, Céline Orvain, Kiran Raosaheb Patil, John Penders, Søren Persson, Nicolas Pons, Milena Popova, Anne Salonen, Delphine Saulnier, Karen P. Scott, Bhagirath Singh, Kathleen Slezak, Patrick Veiga, James Versalovic, Liping Zhao, Erwin G. Zoetendal, S. Dusko Ehrlich, Joel Dore, and Peer Bork (Nov. 2017). "Towards standards for human fecal sample processing in metagenomic studies". In: *Nature Biotechnology* 35.11, pp. 1069–1076. ISSN: 1546-1696. DOI: 10.1038/nbt.3960. URL: https://www.nature.com/articles/nbt.3960 (visited on 05/22/2026).

Debelius, Justine, Se Jin Song, Yoshiki Vazquez-Baeza, Zhenjiang Zech Xu, Antonio Gonzalez, and Rob Knight (Oct. 19, 2016). "Tiny microbes, enormous impacts: what matters in gut microbiome studies?" In: *Genome Biology* 17.1, p. 217. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1086-x. URL: https://doi.org/10.1186/s13059-016-1086-x (visited on 06/06/2026).

Duncan, Anthony, Wing Koon, Katarzyna Sidorczuk, Christopher Quince, Clémence Frioux, and Falk Hildebrand (Nov. 26, 2025). "Detecting signatures underlying the composition of biological data". In: *Nucleic Acids Research* 53.22, gkaf1388. ISSN: 1362-4962. DOI: 10.1093/nar/gkaf1388.

Egger, Roman and Joanne Yu (May 6, 2022). "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts". In: *Frontiers in Sociology* 7. ISSN: 2297-7775. DOI: 10.3389/fsoc.2022.886498. URL: https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full (visited on 05/20/2026).

Fan, Yong and Oluf Pedersen (Jan. 2021). "Gut microbiota in human metabolic health and disease". In: *Nature Reviews Microbiology* 19.1, pp. 55–71. ISSN: 1740-1534. DOI: 10.1038/s41579-020-0433-9. URL: https://www.nature.com/articles/s41579-020-0433-9 (visited on 05/22/2026).

Ferrari, Silvia and Francisco Cribari-Neto (Aug. 1, 2004). "Beta Regression for Modelling Rates and Proportions". In: *Journal of Applied Statistics* 31.7. _eprint: https://doi.org/10.1080/0266476 pp. 799–815. ISSN: 0266-4763. DOI: 10.1080/0266476042000214501. URL: https://doi.org/10.1080/0266476042000214501 (visited on 01/16/2026).

Forbes, Catherine, Merran Evans, Nicholas Hastings 1937-...., and Brian Peacock (2011). *Statistical distributions*. 4th ed. Hoboken: Wiley. xviii, 212 p. ISBN: 978-0-470-39063-4. URL: https://ils.bib.uclouvain.be/global/documents/1677857.

Forbes, Catherine, Merran Evans, Nicholas Hastings, and Brian Peacock (2010a). "Beta Distribution". In: *Statistical Distributions*. Section: 8 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1 John Wiley & Sons, Ltd, pp. 55–61. ISBN: 978-0-470-62724-2. DOI: 10.1002/9780470627242.ch8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470627242.ch8 (visited on 06/02/2026).

– (2010b). "Dirichlet Distribution". In: *Statistical Distributions*. Section: 13 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470627242.ch13. John Wiley & Sons, Ltd, pp. 77–78. ISBN: 978-0-470-62724-2. DOI: 10.1002/9780470627242.ch13. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470627242.ch13 (visited on 06/02/2026).

France, Michael T., Bing Ma, Pawel Gajer, Sarah Brown, Michael S. Humphrys, Johanna B. Holm, L. Elaine Waetjen, Rebecca M. Brotman, and Jacques Ravel (Dec. 2020). "VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition". In: *Microbiome* 8.1, p. 166. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00934-6. URL: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00934-6 (visited on 08/02/2021).

Fukuyama, Julia, Kris Sankaran, and Laura Symul (Sept. 12, 2021). "Multiscale Analysis of Count Data through Topic Alignment". In: *arXiv:2109.05541 [stat]*. arXiv: 2109.05541. URL: http://arxiv.org/abs/2109.05541 (visited on 09/15/2021).

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue (Nov. 15, 2017). "Microbiome Datasets Are Compositional: And This Is Not Optional". In: *Frontiers in Microbiology* 8. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02224.

URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.02224/full (visited on 03/20/2026).

Grün, Bettina and Kurt Hornik (2011). "**topicmodels** : An *R* Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13. ISSN: 1548-7660. DOI: 10.18637/jss.v040.i13. URL: http://www.jstatsoft.org/v40/i13/ (visited on 08/10/2021).

Hawinkel, Stijn, Federico Mattiello, Luc Bijnens, and Olivier Thas (Jan. 18, 2019). "A broken promise: microbiome differential abundance methods do not control the false discovery rate". In: *Briefings in Bioinformatics* 20.1, pp. 210–221. ISSN: 1477-4054. DOI: 10.1093/bib/bbx104. URL: https://doi.org/10.1093/bib/bbx104 (visited on 05/21/2026).

Holmes, Ian, Keith Harris, and Christopher Quince (Feb. 3, 2012). "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics". In: *PLOS ONE* 7.2, e30126. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0030126. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030126 (visited on 12/02/2025).

Hornik, Kurt (Sept. 20, 2005). "A CLUE for CLUster Ensembles". In: *Journal of Statistical Software* 14, pp. 1–25. ISSN: 1548-7660. DOI: 10.18637/jss.v014.i12. URL: https://doi.org/10.18637/jss.v014.i12 (visited on 04/25/2026).

Hosoda, Shion, Suguru Nishijima, Tsukasa Fukunaga, Masahira Hattori, and Michiaki Hamada (June 23, 2020). "Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation". In: *Microbiome* 8.1, p. 95. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00864-3. URL: https://doi.org/10.1186/s40168-020-00864-3 (visited on 12/02/2025).

Hou, Kaijian, Zhuo-Xun Wu, Xuan-Yu Chen, Jing-Quan Wang, Dongya Zhang, Chuanxing Xiao, Dan Zhu, Jagadish B. Koya, Liuya Wei, Jilin Li, and Zhe-Sheng Chen (Apr. 23, 2022). "Microbiota in health and diseases". In: *Signal Transduction and Targeted Therapy* 7.1, p. 135. ISSN: 2059-3635. DOI: 10.1038/s41392-022-00974-4. URL: https://www.nature.com/articles/s41392-022-00974-4 (visited on 11/22/2025).

Huo, Peiyang, Pablo Vargas Ribera, Hans Rediers, and Jan Aerts (Nov. 23, 2025). "Latent Dirichlet Allocation reveals tomato root-associated bacterial interactions responding to hairy root disease". In: *Environmental Microbiome* 20.1, p. 161. ISSN: 2524-6372. DOI: 10.1186/s40793-025-00822-2. URL: https://doi.org/10.1186/s40793-025-00822-2 (visited on 06/03/2026).

Kaul, Abhishek, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada (Nov. 7, 2017). "Analysis of Microbiome Data in the Presence of Excess Zeros". In: *Frontiers in Microbiology* 8. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02114. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.02114/full (visited on 05/21/2026).

Kelly, Brendan J., Robert Gross, Kyle Bittinger, Scott Sherrill-Mix, James D. Lewis, Ronald G. Collman, Frederic D. Bushman, and Hongzhe Li (Aug. 1, 2015). "Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA". In: *Bioinformatics* 31.15, pp. 2461–2468. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv183.

Knight, Rob, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, Alexey V. Melnik, James T. Morton, Jose Navas, Robert A. Quinn, Jon G. Sanders, Austin D. Swafford, Luke R. Thompson, Anupriya Tripathi, Zhenjiang Z. Xu, Jesse R. Zaneveld, Qiyun Zhu, J. Gregory Caporaso, and Pieter C. Dorrestein (July

2018). “Best practices for analysing microbiomes”. In: *Nature Reviews Microbiology* 16.7, pp. 410–422. ISSN: 1740-1534. DOI: 10.1038/s41579-018-0029-9. URL: https://www.nature.com/articles/s41579-018-0029-9 (visited on 05/22/2026).

Koren, Omry, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E. Ley (Jan. 10, 2013). “A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets”. In: *PLOS Computational Biology* 9.1, e1002863. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002863. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002863 (visited on 05/21/2026).

Kuhn, H. W. (2005). “The Hungarian method for the assignment problem”. In: *Naval Research Logistics (NRL)* 52.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.20053, pp. 7–21. ISSN: 1520-6750. DOI: 10.1002/nav.20053. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.20053 (visited on 04/25/2026).

Lederberg, Joshua and Alexa T McCray (2001). “’Ome Sweet ’Omics– A Genealogical Treasury of Words”. In.

Li, Hongzhe (2015). “Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis”. In: *Annual Review of Statistics and Its Application*. DOI: 10.1146/annurev-statistics-010814-020351. URL: https://sci-hub.box/10.1146/annurev-statistics-010814-020351 (visited on 06/02/2026).

Liu, Xiaoni (Sept. 30, 2016). “Microbiome”. In: *The Yale Journal of Biology and Medicine* 89.3, pp. 275–276. ISSN: 0044-0086. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5045136/ (visited on 06/02/2026).

Lozupone, Catherine A., Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight (Sept. 2012). “Diversity, stability and resilience of the human gut microbiota”. In: *Nature* 489.7415, pp. 220–230. ISSN: 1476-4687. DOI: 10.1038/nature11550. URL: https://www.nature.com/articles/nature11550 (visited on 12/11/2025).

Ma, Siyuan, Boyu Ren, Himel Mallick, Yo Sup Moon, Emma Schwager, Sagun Maharjan, Timothy L. Tickle, Yiren Lu, Rachel N. Carmody, Eric A. Franzosa, Lucas Janson, and Curtis Huttenhower (Sept. 13, 2021). “A statistical model for describing and simulating microbial community profiles”. In: *PLOS Computational Biology* 17.9, e1008913. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008913. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008913 (visited on 04/24/2026).

Manichanh, Chaysavanh, Natalia Borruel, Francesc Casellas, and Francisco Guarner (Oct. 2012). “The gut microbiota in IBD”. In: *Nature Reviews Gastroenterology & Hepatology* 9.10, pp. 599–608. ISSN: 1759-5053. DOI: 10.1038/nrgastro.2012.152. URL: https://www.nature.com/articles/nrgastro.2012.152 (visited on 05/22/2026).

McLaren, Michael R, Amy D Willis, and Benjamin Callahan (Sept. 10, 2019). “Consistent and correctable bias in metagenomic sequencing experiments”. In: *eLife* 8. Ed. by Peter Turnbaugh, Wendy S Garrett, Peter Turnbaugh, Christopher Quince, and Sean Gibbons, e46923. ISSN: 2050-084X. DOI: 10.7554/eLife.46923. URL: https://doi.org/10.7554/eLife.46923 (visited on 05/22/2026).

McMurdie, Paul J. and Susan Holmes (Apr. 3, 2014). “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible”. In: *PLOS Computational Biology* 10.4, e1003531. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003531. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531 (visited on 05/22/2026).

Mimno, David and Andrew McCallum (June 13, 2012). *Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression*. DOI: 10.48550/arXiv.1206.3278. arXiv: 1206.3278[cs]. URL: http://arxiv.org/abs/1206.3278 (visited on 02/07/2026).

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazi Evangelista, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Cameron Martino, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J.F. Ter Braak, and James Weedon (May 25, 2026). "vegan: Community Ecology Package". In: Institution: Comprehensive R Archive Network Pages: 2.7-5. DOI: 10.32614/CRAN.package.vegan. URL: https://CRAN.R-project.org/package=vegan (visited on 06/04/2026).

Opal, Steven M. (2010). "A Brief History of Microbiology and Immunology". In: *Vaccines: A Biography*. Ed. by Andrew W. Artenstein. New York, NY: Springer, pp. 31–56. ISBN: 978-1-4419-1108-7. DOI: 10.1007/978-1-4419-1108-7_3. URL: https://doi.org/10.1007/978-1-4419-1108-7_3 (visited on 12/06/2025).

Pappalardo, Vincent Y., Leyla Azarang, Egija Zaura, Bernd W. Brandt, and Renée X. de Menezes (Feb. 5, 2024). "A new approach to describe the taxonomic structure of microbiome and its application to assess the relationship between microbial niches". In: *BMC Bioinformatics* 25.1, p. 58. ISSN: 1471-2105. DOI: 10.1186/s12859-023-05575-8. URL: https://doi.org/10.1186/s12859-023-05575-8 (visited on 06/03/2026).

Paulson, Joseph N., O. Colin Stine, Héctor Corrada Bravo, and Mihai Pop (Dec. 2013). "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12, pp. 1200–1202. ISSN: 1548-7091. DOI: 10.1038/nmeth.2658. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4010126/ (visited on 05/21/2026).

Ravel, J., P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, R. M. Brotman, C. C. Davis, K. Ault, L. Peralta, and L. J. Forney (Mar. 15, 2011). "Vaginal microbiome of reproductive-age women". In: *Proceedings of the National Academy of Sciences* 108 (Supplement_1), pp. 4680–4687. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1002611107. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1002611107 (visited on 08/02/2021).

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley (2019). "**stm** : An *R* Package for Structural Topic Models". In: *Journal of Statistical Software* 91.2. ISSN: 1548-7660. DOI: 10.18637/jss.v091.i02. URL: http://www.jstatsoft.org/v91/i02/ (visited on 09/12/2025).

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (Oct. 2014). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4, pp. 1064–1082. ISSN: 0092-5853, 1540-5907. DOI: 10.1111/ajps.12103. URL: https://onlinelibrary.wiley.com/doi/10.1111/ajps.12103 (visited on 09/12/2025).

Robinson, Courtney K., Rebecca M. Brotman, and Jacques Ravel (May 1, 2016). "Intricacies of assessing the human microbiome in epidemiologic studies". In: *Annals of Epidemiology*. The Microbiome and Epidemiology 26.5, pp. 311–321. ISSN: 1047-2797. DOI: 10.1016/j.annepidem.2016.04.005. URL: https://www.sciencedirect.com/science/article/pii/S1047279716300874 (visited on 05/21/2026).

Sankaran, Kris and Susan Holmes (Oct. 1, 2019). "Latent variable modeling for the microbiome". In: *Biostatistics* 20.4, pp. 599–614. ISSN: 1465-4644, 1468-4357. DOI: 10.1093/biostatistics/kxy018. URL: https://academic.oup.com/biostatistics/article/20/4/599/5032578 (visited on 08/10/2021).

Sender, Ron, Shai Fuchs, and Ron Milo (Jan. 28, 2016). "Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans". In: *Cell* 164.3, pp. 337–340. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.01.013. URL: https://www.sciencedirect.com/science/article/pii/S0092867416000532 (visited on 11/22/2025).

Symul, Laura, Pratheepa Jeganathan, Elizabeth K. Costello, Michael France, Seth M. Bloom, Douglas S. Kwon, Jacques Ravel, David A. Relman, and Susan Holmes (Nov. 29, 2023). "Sub-communities of the vaginal microbiota in pregnant and non-pregnant women". In: *Proceedings of the Royal Society B: Biological Sciences* 290.2011, p. 20231461. DOI: 10.1098/rspb.2023.1461. URL: https://royalsocietypublishing.org/doi/10.1098/rspb.2023.1461 (visited on 01/12/2025).

Thorsen, Jonathan, Asker Brejnrod, Martin Mortensen, Morten A. Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage (Nov. 25, 2016). "Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies". In: *Microbiome* 4.1, p. 62. ISSN: 2049-2618. DOI: 10.1186/s40168-016-0208-8. URL: https://doi.org/10.1186/s40168-016-0208-8 (visited on 05/21/2026).

Tourlousse, Dieter M., Satowa Yoshiike, Akiko Ohashi, Satoko Matsukura, Naohiro Noda, and Yuji Sekiguchi (Feb. 28, 2017). "Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing". In: *Nucleic Acids Research* 45.4, e23. ISSN: 1362-4962. DOI: 10.1093/nar/gkw984.

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon (Oct. 2007). "The Human Microbiome Project". In: *Nature* 449.7164, pp. 804–810. ISSN: 1476-4687. DOI: 10.1038/nature06244. URL: https://www.nature.com/articles/nature06244 (visited on 06/02/2026).

Valle, Denis, Gilson Shimizu, Rafael Izbicki, Leandro Maracahipes, Divino Vicente Silverio, Lucas N. Paolucci, Yusuf Jameel, and Paulo Brando (2021). "The Latent Dirichlet Allocation model with covariates (LDAcov): A case study on the effect of fire on species composition in Amazonian forests". In: *Ecology and Evolution* 11.12. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.7626, pp. 7970–7979. ISSN: 2045-7758. DOI: 10.1002/ece3.7626. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.7626 (visited on 02/05/2026).

Vandeputte, Doris, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y. Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony, and Jeroen Raes (Nov. 23, 2017). "Quantitative microbiome profiling links gut community variation to microbial load". In: *Nature* 551.7681, pp. 507–511. ISSN: 1476-4687. DOI: 10.1038/nature24460.

Wallach, Hanna, David Mimno, and Andrew McCallum (2009). "Rethinking LDA: Why Priors Matter". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta. Vol. 22. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf.

Wang, Yu-Xiong and Yu-Jin Zhang (June 2013). "Nonnegative Matrix Factorization: A Comprehensive Review". In: *IEEE Transactions on Knowledge and Data Engineering*

25.6, pp. 1336–1353. ISSN: 1558-2191. DOI: 10.1109/TKDE.2012.51. URL: https://ieeexplore.ieee.org/document/6165290/ (visited on 05/20/2026).

Wesolowska-Andersen, Agata, Martin Iain Bahl, Vera Carvalho, Karsten Kristiansen, Thomas Sicheritz-Pontén, Ramneek Gupta, and Tine Rask Licht (June 5, 2014). "Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis". In: *Microbiome* 2.1, p. 19. ISSN: 2049-2618. DOI: 10.1186/2049-2618-2-19. URL: https://doi.org/10.1186/2049-2618-2-19 (visited on 06/06/2026).

Willis, Amy D. (Oct. 23, 2019). "Rarefaction, Alpha Diversity, and Statistics". In: *Frontiers in Microbiology* 10. ISSN: 1664-302X. DOI: 10.3389/fmicb.2019.02407. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.02407/full (visited on 06/04/2026).

Yatsunenko, Tanya, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon (June 2012). "Human gut microbiome viewed across age and geography". In: *Nature* 486.7402, pp. 222–227. ISSN: 1476-4687. DOI: 10.1038/nature11053. URL: https://www.nature.com/articles/nature11053 (visited on 05/22/2026).

Zhernakova, Alexandra, Alexander Kurilshikov, Marc Jan Bonder, Ettje F. Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, Jun Wang, Floris Imhann, Eelke Brandsma, Soesma A. Jankipersadsing, Marie Joossens, Maria Carmen Cenit, Patrick Deelen, Morris A. Swertz, LifeLines cohort study, Rinse K. Weersma, Edith J. M. Feskens, Mihai G. Netea, Dirk Gevers, Daisy Jonkers, Lude Franke, Yurii S. Aulchenko, Curtis Huttenhower, Jeroen Raes, Marten H. Hofker, Ramnik J. Xavier, Cisca Wijmenga, and Jingyuan Fu (Apr. 29, 2016). "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity". In: *Science* 352.6285, pp. 565–569. DOI: 10.1126/science.aad3369. URL: https://www.science.org/doi/10.1126/science.aad3369 (visited on 06/06/2026).

Inferring cohort differences in microbiota composition using Latent Dirichlet allocation model

Maxime Cornez

Microbiota composition data are inherently characterized by high sparsity and stochastic overdispersion, complicating the reliable detection of cohort-level differences. To address this challenge, this thesis evaluates the capacity of a joint analytical framework, combining Latent Dirichlet Allocation (LDA) and PERMANOVA, to infer and test global differences between microbiota cohorts. An in silico generative process coupled with a full factorial design is implemented to evaluate the pipeline against controlled variations in biological signal amplitude (A), cohort size (N_g), and the Negative Binomial distribution size parameter (ϕ). By exclusively simulating the cohort effect on the latent topic prevalence matrix (Γ), this framework isolates the statistical behavior of the pipeline from the confounding factors of empirical datasets. The systematic evaluation reveals a crucial functional decoupling between latent space reconstruction and statistical testing. First, the structural fidelity of the LDA inference is primarily dictated by overdispersion; extreme variance ($\phi \leq 1$) degrades the accuracy of the inferred matrices, a degradation that increased cohort recruitment alone cannot fully compensate for. Despite this, the downstream statistical testing phase demonstrates consistent reliability. The framework maintains a controlled Type I error rate (oscillating around 5%), ensuring that extreme overdispersion does not translate into spurious cohort separations. Furthermore, the statistical power is primarily governed by the biological effect amplitude and cohort size, exhibiting compensatory dynamics that facilitate accurate detection provided the biological contrast is pronounced or sampling is sufficient. Consequently, this thesis quantitatively maps the operational limits of the LDA-PERMANOVA pipeline, providing data-driven guidelines for the experimental design of future microbiota studies.

UNIVERSITÉ CATHOLIQUE de LOUVAIN

Faculté des bioingénieurs

Croix du Sud, 2bte L7.05.01, 1348 Louvain-La-Neuve, Belgique | **www.uclouvain.be/agro**