Statistica Sinica Preprint No: SS-2026-0149	
Title	Differential Network Analysis for Compositional Data
Manuscript ID	SS-2026-0149
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0149
Complete List of Authors	Chencheng Ma, Hao Chen, Yijuan Hu and Xiang Zhan
Corresponding Authors	Xiang Zhan
E-mails	zhanx@seu.edu.cn
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Differential Network Analysis for Compositional Data

Chencheng Ma[1], Hao Chen[1], Yijuan Hu[2] and Xiang Zhan[3]

[1]*Shandong University,* [2]*Peking University,* [3]*Southeast University*

Abstract: It is often of interest to understand how conditional dependence relationships among components of a compositional vector differs between two conditions. In this paper, we specify such a conditional dependence network among components using a compositional precision matrix. To perform differential network analysis for compositional data, we propose DNA-Coda that directly estimates the difference of two compositional precision matrices. Under mild regularity conditions, our approach is shown to be consistent in support recovery and estimation. Comprehensive simulations demonstrate DNA-Coda's statistical superiority over existing methods in the accuracy of estimation and recovery of true differential network matrix. The potential usefulness of our method is also illustrated by an application example to differential network analysis of microbiome compositional data.

Key words and phrases: Compositional data analysis (Coda), Differential network analysis, High dimensionality, Precision matrix

1. Introduction

Compositional data arise in many fields, including ecology, the social sciences, and biomedical research. Their components describe relative contributions to a whole and are commonly represented by proportions that sum to one. A prominent example is 16S rRNA gene sequencing, for which microbial sequencing counts are often converted to relative abundances

The correspondence should be addressed to Xiang Zhan (zhanx@seu.edu.cn), Yijuan Hu(yijuanhu@bicmr.pku.edu.cn) or Hao Chen (hao.chen@sdu.edu.cn)

by dividing each taxon's count by the sample total (Fernandes et al., 2014). Because these proportions have a fixed sum, a change in the relative abundance of one taxon necessarily changes the relative contributions of other taxa, even when the corresponding unnormalized absolute abundances do not exhibit the same association. This constraint, known as the *closure effect*, complicates the interpretation of correlations computed directly from compositional observations (Aitchison, 1982). It therefore motivates methods that distinguish dependence patterns in the observed compositions from those of the underlying variables.

In this article, we consider such a compositional data analysis (Coda) task under the setting of differential analysis, with the aim of identifying changes in association structure between two conditions. Much of the existing literature on differential analysis of compositions has focused on changes in the relative abundances of individual components (Cao et al., 2018; Brill et al., 2022; Zhou et al., 2022). However, two conditions may have similar marginal abundance patterns while exhibiting different associations among components (Bandyopadhyay et al., 2010; Jiang et al., 2021). Precision matrices offer a way to represent partial associations among variables; under a multivariate Gaussian model, their zero entries additionally characterize conditional independence. We focus on the difference between the precision matrices of two latent log-basis vectors, namely, the componentwise logarithms of the underlying positive, unnormalized absolute abundances. This differential precision matrix is our target, rather than the difference between networks constructed directly from observed relative abundances.

Early methodological advances in Coda centered on the use of log-ratio transformations to project data from the simplex into a Euclidean space, thereby mitigating the pervasive closure effect (Aitchison, 2003). In particular, the centered log-ratio (CLR) transformation

subtracts the mean log relative abundance from each log component, producing a vector whose entries sum to zero. The resulting covariance matrix is singular, but its Moore–Penrose inverse defines a population-level compositional precision matrix. Compositional observations retain relative abundance information but not the overall absolute abundance scale. Consequently, the differential precision matrix defined on the CLR scale generally differs from the corresponding differential precision matrix of the latent log-basis variables. In the differential setting, this raises a fundamental question: under what structural conditions can the difference between two CLR-scale compositional precision matrices serve as an good proxy for the difference between their latent basis counterparts? Answering this question is necessary for establishing statistical guarantees for estimation of the latent differential network from compositional observations.

Several methods have been developed for precision matrix estimation from compositional data (Kurtz et al., 2015; Fang et al., 2017; Fang, 2023; Zhang et al., 2025). In particular, the CARE framework (Zhang et al., 2025) establishes approximate identifiability and high-dimensional statistical guarantees for estimating a single sparse latent basis precision matrix. Differential network analysis, however, presents a distinct statistical problem: the target is the difference between two latent basis precision matrices, which may be sparse even when the individual matrices are dense. Consequently, separately estimating the two precision matrices using sparse individual-network estimators may impose unnecessarily restrictive structural assumptions. Although direct differential network estimators for compositional data have been proposed (He and Deng, 2019), establishing theoretical guarantees for the latent differential precision matrix remains challenging because its CLR-scale counterpart does not generally coincide with the latent target. A central question is whether this dis-

crepancy can be controlled under structural conditions on the differential matrix itself, without requiring sparsity of the individual precision matrices. Motivated by this problem, we develop and analyze **DNA-Coda** (**D**ifferential **N**etwork **A**nalysis for **Co**mpositional **da**ta), a theoretically justified framework for direct differential precision-matrix estimation from compositional observations. Our main contributions are as follows.

1. **Approximate identifiability of the latent differential network.** We establish an explicit elementwise bound between $\mathbf{\Delta}_0$ and $\mathbf{\Delta}_c$. Under uniform spectral bounds on the individual latent precision matrices and a fixed elementwise ℓ_1 bound on their difference, the identification error is of order $O(p^{-1/2})$.

2. **Convergence rates and support recovery guarantees.** Under differential sparsity, a coherence condition, and either sub-Gaussian or polynomial-type tail assumptions, we derive convergence rates for the DNA-Coda estimator in the elementwise maximum and Frobenius norms. The resulting bounds distinguish sampling estimation error from the identification error due to compositional observations. Under a minimum-signal condition, we further establish a signed-support recovery guarantee for a thresholded estimator, without imposing sparsity on the two individual latent precision matrices.

3. **Implementation and numerical investigation.** Building on the symmetric constrained ℓ_1-minimization formulation and alternating direction method of multipliers (ADMM) strategy of Zhao et al. (2014), we implement a direct estimator using CLR sample covariance matrices. Simulations compare DNA-Coda with its oracle counterpart based on latent basis vectors, a separate-estimation approach based on CARE (CARE-S) and its oracle version (CARE-S-Oracle), and DCDTr, an existing direct

Statistica Sinica: Newly accepted Paper

compositional differential-network estimator. An application to mucosal microbiome data illustrates the differential association patterns identified between familial adenomatous polyposis (FAP) and non-FAP groups.

1.1 Related Work

The most directly related compositional differential-network study is He and Deng (2019). They proposed DCDTr, which incorporates CLR covariance information into a lasso-penalized D-trace loss to estimate differential networks directly, rather than first estimating two individual precision matrices. They also considered a compositional adaptation of the constrained ℓ_1-minimization approach of Zhao et al. (2014), which is closely related to our proposed estimator. Furthermore, He and Deng (2019) discussed approximating the latent log-basis covariance matrices by their CLR-scale counterparts as an alternative motivation for the DCDTr loss and as a basis for other differential-network estimators. However, their derivation of DCDTr's population minimization property relies on an exchangeability condition, and they did not establish an explicit approximation bound between the latent basis and CLR-scale differential precision matrices, or high-dimensional convergence rates and support-recovery guarantees for the latent differential target. Our emphasis is therefore on establishing approximate identifiability and rigorous statistical guarantees for the latent differential precision matrix under structural conditions imposed on the difference, rather than on the individual precision matrices.

A second line of research concerns precision matrix estimation within a single compositional population (Kurtz et al., 2015; Fang et al., 2017; Fang, 2023). Of particular relevance, the Composition Adaptive Regularized Estimation (CARE) framework of Zhang et al. (2025)

Statistica Sinica: Newly accepted Paper

relates the compositional precision matrix to its latent log-basis counterpart and establishes statistical guarantees for estimating a sparse basis precision matrix. Our setting differs in its target and structural assumptions: we estimate the difference between two latent precision matrices and do not require either individual matrix to be sparse. The identification and estimation errors must therefore be controlled at the level of the differential matrix.

For ordinary Euclidean observations, related methods include joint precision-matrix estimation, such as the fused graphical lasso (Danaher et al., 2014), and direct differential precision-matrix estimation based on constrained ℓ_1-minimization or D-trace losses (Zhao et al., 2014; Yuan et al., 2017). These methods provide important methodological foundations but do not, by themselves, address the identification gap introduced by compositional observations. From a different inferential perspective, Cai et al. (2019) developed a Markov random field approach with global and multiple-testing procedures for detecting differential microbial networks. Their testing objective is distinct from direct estimation of the latent log-basis differential precision matrix considered here.

1.2 Organization of the Article

The remainder of the paper is organized as follows. We first introduce the notation and describe the proposed method in Section 2. Section 3 establishes theoretical guarantees for our DNA-Coda estimator in terms of both rates of convergence and support recovery. Section 4 validates the method through simulations, and Section 5 applies it to a real-world microbiome dataset. The final section discusses the findings and limitations. The proofs and implementation details of the proposed methodology are provided in the online supplementary material.

2. Differential Network Analysis for Compositional Data

Notations: We use $|\cdot|$ to denote element-wise norms and $\|\cdot\|$ to denote matrix norms. For a $p \times 1$ vector $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$, define $|\mathbf{a}|_0$ to be the number of nonzero elements of $\mathbf{a}$, and let $|\mathbf{a}|_1 = \sum_j |a_j|$, $|\mathbf{a}|_2 = \left(\sum_j a_j^2\right)^{1/2}$ and $|\mathbf{a}|_\infty = \max_j |a_j|$. For a $p \times p$ matrix $\mathbf{A}$ with entries a_{jk}, let $|\mathbf{A}|_0$ be the number of nonzero entries of $\mathbf{A}$, $|\mathbf{A}|_1 = \sum_{j,k} |a_{jk}|$, $|\mathbf{A}|_\infty = \max_{j,k} |a_{jk}|$, $\|\mathbf{A}\|_1 = \max_k \sum_j |a_{jk}|$, $\|\mathbf{A}\|_\infty = \max_j \sum_k |a_{jk}|$, $\|\mathbf{A}\|_2 = \sup_{|\mathbf{a}|_2 \leqslant 1} |\mathbf{A}\mathbf{a}|_2$ and $\|\mathbf{A}\|_{\mathrm{F}} = \left(\sum_{j,k} a_{jk}^2\right)^{1/2}$. Also, $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of $\mathbf{A}$, respectively, and $\mathbf{A} \succ 0$ stands for $\mathbf{A}$ being symmetric and positive definite. Finally, let $\mathbf{I}_p$ denote the $p \times p$ identity matrix and $\mathbf{1}_p$ be the p-vector of ones.

2.1 Compositional Differential Network Matrix

Consider an observed compositional vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ that takes values in the $(p-1)-$dimensional unit simplex $\mathcal{S}^{p-1} = \left\{(x_1, \ldots, x_p) : x_j > 0 \text{ for } j = 1, \ldots, p, \sum_{j=1}^p x_j = 1\right\}$. Let $\mathbf{W}_X = (W_{X,1}, \ldots, W_{X,p})^T$ with $W_{X,j} > 0$ for all j be a latent vector, called the *basis*, which is typically unobservable but can generate the observed compositions. That is, the observed compositions $(X_1, \ldots, X_p)$ can be viewed as a realization of

$$\left(\frac{W_{X,1}}{\sum_{j=1}^p W_{X,j}}, \ldots, \frac{W_{X,p}}{\sum_{j=1}^p W_{X,j}}\right).$$

To specify a basis covariance structure compatible with the log-ratio form of compositional covariance structure (described later in the next paragraph), we define the log-basis vector $\mathbf{T}_X = (T_{X,1}, \ldots, T_{X,p})^T$ with $T_{X,j} = \log\{W_{X,j}\}$ for all j. Let $\mathbf{\Omega}_{X,0} = \left(\omega^0_{X,ij}\right)$ denote the precision matrix of $\mathbf{T}_X$, and $\mathbf{\Sigma}_{X,0} = \left(\sigma^0_{X,ij}\right) = \mathbf{\Omega}^{-1}_{X,0}$ denote the covariance matrix of $\mathbf{T}_X$. $\mathbf{\Sigma}_{X,0}$ and $\mathbf{\Omega}_{X,0}$ are called basis covariance matrix and basis precision matrix, respectively, hereafter for ease of reference.

Statistica Sinica: Newly accepted Paper

One common way in Coda is via log-ratio transformations that move data from simplex to reals to remove the compositional effect in components. In particular, we can apply the centered log-ratio (CLR) transformation (Aitchison, 2003) to calculate $\mathbf{Z}_X = (Z_{X,1}, \ldots, Z_{X,p})^T$ with $Z_{X,j} = \log\{W_{X,j}/g(\mathbf{W_X})\}$, where $g(\mathbf{W_X}) = \left(\prod_{j=1}^{p} W_{X,j}\right)^{1/p}$ is the geometric mean function. Due to the scale invariance of the CLR transformation, we have

$$\frac{W_{X,j}}{g(\mathbf{W}_X)} = \frac{W_{X,j}/\sum_l W_{X,l}}{g(\mathbf{W}_X/\sum_l W_{X,l})}, \quad j = 1, \ldots, p,$$

where $W_{X,j}/\sum_l W_{X,l}$ can be approximated by its sample version X_j. That is, the CLR-transformed vector $\mathbf{Z}_X$ can be directly estimated by the observed compositional vector $\mathbf{X}$, without requiring knowledge of the latent basis vector $\mathbf{W}_X$. Accordingly, the CLR covariance matrix $\mathbf{\Sigma}_{X,c} = \left(\sigma^c_{X,ij}\right)$ of $\mathbf{Z}_X$ can be estimated from the observed compositions. To distinguish it from the basis covariance matrix $\mathbf{\Sigma}_{X,0}$, we call $\mathbf{\Sigma}_{X,c}$ the compositional covariance matrix hereafter.

Owing to the zero-sum constraint on components of $\mathbf{Z}_X$, $\mathbf{\Sigma}_{X,c}$ has an eigenvalue zero and the corresponding eigenvector $\mathbf{v}_{X,p} = \mathbf{1}_p/\sqrt{p}$. Let $\mathbf{\Sigma}_{X,c} = \mathbf{V}_X\mathbf{D}_{X,c}\mathbf{V}_X^T$ be the spectral decomposition of $\mathbf{\Sigma}_{X,c}$. The rank-deficient matrix $\mathbf{\Sigma}_{X,c}$ can be made invertible by a rank-1 correction. That is, we replace the only zero eigenvalue with any positive constant $\rho_X > 0$, and correspondingly, $\mathbf{\Sigma}_{X,c} + \rho_X\mathbf{v}_{X,p}\mathbf{v}_{X,p}^T$ is nonsingular. We now define the compositional precision matrix $\mathbf{\Omega}_{X,c}$ as the Moore–Penrose inverse $\mathbf{\Sigma}^+_{X,c}$ of the compositional covariance matrix $\mathbf{\Sigma}_{X,c}$. That is,

$$\mathbf{\Omega}_{X,c} := \mathbf{\Sigma}^+_{X,c} = \mathbf{V}_X\mathbf{D}^+_{X,c}\mathbf{V}_X^T = \left(\mathbf{\Sigma}_{X,c} + \rho_X\mathbf{v}_{X,p}\mathbf{v}_{X,p}^T\right)^{-1} - \frac{1}{\rho_X}\mathbf{v}_{X,p}\mathbf{v}_{X,p}^T. \tag{2.1}$$

Finally, the following proposition is useful for calculations involving $\mathbf{\Omega}_{X,c}$.

Statistica Sinica: Newly accepted Paper

Proposition 1. *Compositional precision matrix* $\mathbf{\Omega}_{X,c}$ *satisfies the following:*

$$\mathbf{\Omega}_{X,c}\mathbf{\Sigma}_{X,c} = \mathbf{\Sigma}_{X,c}\mathbf{\Omega}_{X,c} = \mathbf{G},$$

$$\mathbf{G}\mathbf{\Omega}_{X,c} = \mathbf{\Omega}_{X,c}\mathbf{G} = \mathbf{\Omega}_{X,c},$$

$$\mathbf{G}\mathbf{\Sigma}_{X,c} = \mathbf{\Sigma}_{X,c}\mathbf{G} = \mathbf{\Sigma}_{X,c},$$

where $\mathbf{G} = \mathbf{I}_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^T$.

Proofs of Proposition 1 are provided in the supplementary material. These algebraic insights established in Proposition 1 will be used to derive estimation equations (2.2) and (2.4), which are essential to develop our constrained ℓ_1-minimization estimator described in the next section.

The aim of this article is to perform differential network analysis of compositional data. To this end, we next introduce the second compositional vector $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$ with analogous notation and properties to $\mathbf{X}$, including associated terms $\mathbf{W}_Y$, $\mathbf{T}_Y$, $\mathbf{Z}_Y$, $\mathbf{\Sigma}_{Y,0}$, $\mathbf{\Sigma}_{Y,c}$, $\mathbf{\Omega}_{Y,0}$, and $\mathbf{\Omega}_{Y,c}$. Then, two differential network matrices are defined as: (1) The *basis differential network matrix* $\mathbf{\Delta}_0 = \mathbf{\Omega}_{Y,0} - \mathbf{\Omega}_{X,0}$, representing the difference in precision matrices on the log-basis scale ($\mathbf{T}_X$, $\mathbf{T}_Y$), which are often unobservable; (2) The *compositional differential network matrix* $\mathbf{\Delta}_c = \mathbf{\Omega}_{Y,c} - \mathbf{\Omega}_{X,c}$, capturing the difference in precision matrices on the CLR scale ($\mathbf{Z}_X$, $\mathbf{Z}_Y$), which can be estimated from observed compositional data.

Given the inherent unobservability of the basis, we must estimate $\mathbf{\Delta}_0$ from the observed compositional data. We first note that Proposition 1 establishes that the true compositional differential network $\mathbf{\Delta}_c$ satisfies

$$\mathbf{\Sigma}_{X,c}\mathbf{\Delta}_c\mathbf{\Sigma}_{Y,c} - (\mathbf{\Sigma}_{X,c} - \mathbf{\Sigma}_{Y,c}) = \mathbf{0}. \tag{2.2}$$

Equation (2.2) is not directly applicable to estimate our target estimand $\mathbf{\Delta}_0$. To bridge this gap, we first establish the approximate identifiability of $\mathbf{\Delta}_0$ through $\mathbf{\Delta}_c$. Following statistical

Statistica Sinica: Newly accepted Paper

literature on high-dimensional precision matrix (Cai et al., 2011, 2016), we assume that latent basis precision matrices $\mathbf{\Omega}_{X,0}$ and $\mathbf{\Omega}_{Y,0}$ belong to the class:

$$\mathcal{U}(M, R_X, R_Y) = \left\{ \begin{array}{l} \mathbf{\Omega}_{X,0} \succ 0,\ 1/R_X \leq \lambda_{\min}(\mathbf{\Omega}_{X,0}) \leq \lambda_{\max}(\mathbf{\Omega}_{X,0}) \leq R_X; \\ \mathbf{\Omega}_{Y,0} \succ 0,\ 1/R_Y \leq \lambda_{\min}(\mathbf{\Omega}_{Y,0}) \leq \lambda_{\max}(\mathbf{\Omega}_{Y,0}) \leq R_Y; \\ |\mathbf{\Delta}_0|_1 \leqslant M \end{array} \right\},$$

where M, $R_X > 1$ and $R_Y > 1$ are constants that are independent of p. This class encompasses a broad family of latent precision matrices. The bounded eigenvalue conditions guarantee that $\mathbf{\Omega}_{X,0}$ and $\mathbf{\Omega}_{Y,0}$ are uniformly well-conditioned, while the element-wise ℓ_1-norm constraint on $\mathbf{\Delta}_0$ regulates the overall scale of their difference. Notably, no sparsity assumption is imposed on the individual precision matrices; the requirement is solely that their discrepancy remains within a controlled magnitude.

Proposition 2. *Assuming $\mathbf{\Omega}_{X,0}, \mathbf{\Omega}_{Y,0} \in \mathcal{U}(M, R_X, R_Y)$, there exists some constant $C_0 > 0$ such that the following bound holds:*

$$|\mathbf{\Delta}_0 - \mathbf{\Delta}_c|_\infty \leq \frac{C_0 M(R_Y^2 + R_Y R_X)}{\sqrt{p}}. \tag{2.3}$$

The proof of Proposition 2 is provided in the supplementary material. Proposition 2 establishes the approximate identifiability of $\mathbf{\Delta}_0$ through $\mathbf{\Delta}_c$, with the identification error bounded by $O(p^{-1/2})$ under the stated conditions. The intuition behind this result is that the CLR transformation induces a correction term for each latent basis precision matrix. The discrepancy between $\mathbf{\Delta}_0$ and $\mathbf{\Delta}_c$ is determined by the difference between these two correction terms, which can be controlled under the bounded-spectrum conditions and the elementwise ℓ_1-norm constraint on $\mathbf{\Delta}_0$. Consequently, the identification error vanishes as $p \to \infty$. Importantly, approximate identifiability does not imply equality of the unthresholded supports of $\mathbf{\Delta}_0$ and $\mathbf{\Delta}_c$. In particular, $\mathbf{\Delta}_c$ may contain small nonzero entries outside

Statistica Sinica: Newly accepted Paper

the support of $\mathbf{\Delta}_0$, even though the maximum-entry discrepancy between the two matrices vanishes with increasing dimension. Therefore, Proposition 2 should be interpreted as an identification result for identifying the latent target rather than as a statement of exact support equivalence. This type of phenomenon implied by Proposition 2 has also been named as the blessing of dimensionality in compositional data analysis (Cao et al., 2019; Zhang et al., 2025). In biological applications, this result provides a theoretical justification for using the CLR-scale differential precision matrix as an approximate population-level proxy for the latent basis differential precision matrix in microbiome studies, where only relative abundance information is typically available.

Remark 1. Proposition 2 is stated under the fixed-M parameter class. To illustrate how the identification error changes when the differential-network complexity increases with p, suppose instead that $|\mathbf{\Delta}_0|_1 \leq M(p)$, while R_X and R_Y remain fixed. By the same argument as in Proposition 2, the identification error is of order $O\{M(p)/\sqrt{p}\}$, and hence $M(p) = o(\sqrt{p})$ is sufficient for the identification error to vanish. If, in addition, the upper-triangular part of $\mathbf{\Delta}_0$, including the diagonal, contains at most $s(p)$ nonzero entries, then the fixed spectral bounds imply

$$|\mathbf{\Delta}_0|_1 \leq 2s(p)(R_X + R_Y),$$

and consequently

$$|\mathbf{\Delta}_0 - \mathbf{\Delta}_c|_\infty = O\left\{\frac{s(p)}{\sqrt{p}}\right\}.$$

Thus, increasing differential-network complexity partially offsets the $p^{-1/2}$ dimensional decay. In particular, the blessing of dimensionality weakens as the number of differential entries grows with p, while $s(p) = o(\sqrt{p})$ remains sufficient for the identification error to vanish.

When compared with Proposition 2 of Zhang et al. (2025), crucial distinctions emerge

Statistica Sinica: Newly accepted Paper

in both the convergence rates and the structure of the bounding constants, reflecting the fundamental difference between estimating a single precision matrix versus a differential network. In terms of rates, CARE achieves a sharper $O(p^{-1})$ rate when the ℓ_1 bound M_p on $\mathbf{\Omega}_0$ is fixed, but requires sparsity of the full basis precision matrix; our $O(p^{-1/2})$ rate is slower but obtained under weaker assumptions, where only $\mathbf{\Delta}_0$ is assumed sparse. In terms of constants, CARE's Proposition 2 requires the ℓ_1-norm bound M_p of the precision matrix to satisfy $M_p = o(\sqrt{p})$ in order to guarantee identifiability. Thus, the constant in their bound implicitly depends on how M_p grows with p. By contrast, our Proposition 2 expresses the error bound directly in terms of the fixed parameters (M, R_X, R_Y), which are independent of p. This formulation makes the role of conditioning explicit and avoids reliance on additional growth assumptions, yielding a more stable and interpretable identifiability guarantee. Thus, although our rate is slower, Proposition 2 provides a broader and practically more relevant identifiability guarantee tailored to the differential network setting. This justifies our strategy of using $\mathbf{\Delta}_c$ as a surrogate for $\mathbf{\Delta}_0$ in developing our estimation procedures in the next section.

2.2 Constrained ℓ_1-minimization for DNA-Coda Estimation

The approximate identifiability established in Proposition 2 motivates us to obtain an estimator of the target estimand $\mathbf{\Delta}_0$ by utilizing the following approximation:

$$\widehat{\mathbf{\Sigma}}_{X,c}\mathbf{\Delta}_0\widehat{\mathbf{\Sigma}}_{Y,c} - (\widehat{\mathbf{\Sigma}}_{X,c} - \widehat{\mathbf{\Sigma}}_{Y,c}) \approx \mathbf{0}. \tag{2.4}$$

As demonstrated later in Lemma 1 in Section 3, (2.4) exhibits asymptotic behavior comparable to the original compositional equation (2.2). In particular, in order to estimate $\mathbf{\Delta}_0$, we need to solve the following convex optimization problem:

$$\text{minimize } |\mathbf{\Delta}|_1 \quad \text{subject to } \left|\widehat{\mathbf{\Sigma}}_{X,c}\mathbf{\Delta}\widehat{\mathbf{\Sigma}}_{Y,c} - \left(\widehat{\mathbf{\Sigma}}_{X,c} - \widehat{\mathbf{\Sigma}}_{Y,c}\right)\right|_\infty \leq \lambda,$$

Statistica Sinica: Newly accepted Paper

and then symmetrizing the solution. Here, tuning parameter $\lambda > 0$ is chosen by cross-validation in practice. This is equivalent to a linear program, as for any three $p \times p$ matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}, \mathrm{vec}(\mathbf{ABC}) = \left(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A}\right) \mathrm{vec}(\mathbf{B})$ where $\otimes$ denotes the Kronecker product and $\mathrm{vec}(\mathbf{B})$ denotes the $p^2 \times 1$ vector obtained by stacking the columns of $\mathbf{B}$. Therefore, $\boldsymbol{\Delta}_0$ could be estimated by solving and then symmetrizing the following objective:

$$\text{minimize } |\boldsymbol{\Delta}|_1 \quad \text{s.t.} \quad \left| \left(\widehat{\boldsymbol{\Sigma}}_{Y,c} \otimes \widehat{\boldsymbol{\Sigma}}_{X,c} \right) \mathrm{vec}(\boldsymbol{\Delta}) - \mathrm{vec}\left(\widehat{\boldsymbol{\Sigma}}_{X,c} - \widehat{\boldsymbol{\Sigma}}_{Y,c} \right) \right|_\infty \leqslant \lambda. \tag{2.5}$$

A detailed algorithm to solve problem (2.5) is described in Section 1 of the supplementary material.

Finally, we describe the cross-validation procedure for choosing the tuning parameter λ. We randomly split the whole dataset into a training set and a test set. To measure the predictive performance of $\widehat{\boldsymbol{\Delta}}$, in a similar spirit to the D-trace loss of Yuan et al. (2017), we consider the loss function

$$L\left(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_{X,c}, \boldsymbol{\Sigma}_{Y,c}\right) = \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Delta}\boldsymbol{\Sigma}_{X,c}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{Y,c}\right) - \mathrm{tr}(\boldsymbol{\Sigma}_{X,c}\boldsymbol{\Delta}) + \mathrm{tr}(\boldsymbol{\Sigma}_{Y,c}\boldsymbol{\Delta}),$$

whose derivative is calculated as:

$$\frac{\partial L}{\partial \boldsymbol{\Delta}} = \boldsymbol{\Sigma}_{X,c}\boldsymbol{\Delta}\boldsymbol{\Sigma}_{Y,c} - \left(\boldsymbol{\Sigma}_{X,c} - \boldsymbol{\Sigma}_{Y,c}\right).$$

For the b-th ($b = 1, \ldots, B$) split, denote by $\widehat{\boldsymbol{\Delta}}^{(1b)}(\lambda)$ the estimator with tuning parameter λ based on the training set, and by $\widehat{\boldsymbol{\Sigma}}_{X,c}^{(2b)}$ and $\widehat{\boldsymbol{\Sigma}}_{Y,c}^{(2b)}$ the sample centered log-ratio covariance matrix based on the test set. We choose the optimal value $\hat{\lambda}$ of λ by minimizing

$$\mathrm{CV}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} L\left\{ \widehat{\boldsymbol{\Delta}}^{(1b)}(\lambda), \widehat{\boldsymbol{\Sigma}}_{X,c}^{(2b)}, \widehat{\boldsymbol{\Sigma}}_{Y,c}^{(2b)} \right\}$$

and compute the estimator $\widehat{\boldsymbol{\Delta}}\left(\hat{\lambda}\right)$ based on the whole dataset.

3. Theoretical Properties

We establish high-dimensional theoretical guarantees for the DNA-Coda estimator $\widehat{\mathbf{\Delta}}$ described in the previous section, focusing on estimation accuracy and support recovery. We begin by introducing the regularity conditions used in the analysis.

Assumption 1. The differential network matrix $\mathbf{\Delta}_0$ contains s nonzero entries in its upper triangular part, where s is a positive constant independent of p.

Assumption 2. Denote the (j,k)th entries of $\mathbf{\Sigma}_{X,c}$ and $\mathbf{\Sigma}_{Y,c}$ as $\sigma_{jk}^{X,c}$ and $\sigma_{jk}^{Y,c}$, respectively, and define:

$$\sigma_{\max}^{X,c} = \max_j \sigma_{jj}^{X,c}, \quad \sigma_{\max}^{Y,c} = \max_j \sigma_{jj}^{Y,c}, \quad \mu_{X,c} = \max_{j\neq k} |\sigma_{jk}^{X,c}|, \quad \mu_{Y,c} = \max_{j\neq k} |\sigma_{jk}^{Y,c}|,$$

$$\sigma_{\min}^{S,c} = \min_{j,k} \left(\sigma_{jj}^{Y,c}\sigma_{jj}^{X,c},\ \sigma_{kk}^{Y,c}\sigma_{jj}^{X,c} + 2\sigma_{kj}^{Y,c}\sigma_{jk}^{X,c} + \sigma_{jj}^{Y,c}\sigma_{kk}^{X,c} \right) > 0,$$

where $\mu_{X,c}$ and $\mu_{Y,c}$ satisfy $\mu_c = 4\max\left(\mu_{X,c}\sigma_{\max}^{Y,c},\ \mu_{Y,c}\sigma_{\max}^{X,c}\right) \leqslant (2s)^{-1}\sigma_{\min}^{S,c}$.

Assumption 1 characterizes a localized differential-network regime in which the changes between the two conditions are concentrated on a controlled number of entries. It is intended to isolate the effect of compositional identifiability from the additional complexity induced by a growing differential support, rather than to assert that the number of differential interactions must remain constant in all applications. As discussed in Remark 1, the blessing of dimensionality weakens as the number of differential entries grows with p. Assumption 2 restricts the compositional covariances between components, ensuring that they are not excessively high from the element-level perspectives. Although it would suffice to require that the largest off-diagonal entry of $\mathbf{S}^{\mathrm{T}}\left(\mathbf{\Sigma}_{Y,c} \otimes \mathbf{\Sigma}_{X,c}\right)\mathbf{S}$ having a magnitude less than $(2s)^{-1}\sigma_{\min}^{S,c}$, Assumption 2 provides a more interpretable condition. Furthermore, we assume that log-basis vectors $\mathbf{T}_X$ and $\mathbf{T}_Y$ are centered at $\mathbf{0}$ and satisfy the following tail conditions:

Assumption 3. (Sub-Gaussian tail) Assume $\log p = o(n_X)$, with constants $\eta > 0$ and $K > 0$ satisfying:

$$\sup_{\mathbf{s}\in\mathbb{R}^p, \|\mathbf{s}\|_2=1} E\left[\exp\left\{\eta\left(\mathbf{s}^T\mathbf{T}_X\right)^2 / \operatorname{Var}\left(\mathbf{s}^T\mathbf{T}_X\right)\right\}\right] \leq K,$$

and $\mathbf{T}_Y$ satisfies an analogous sub-Gaussian tail condition with sample size n_Y.

Assumption 4. (Polynomial-type tail) Assume $p = O(n_X^\gamma)$ for some $\gamma > 0$, with constants $\varepsilon > 0$ and $K' > 0$ satisfying:

$$\sup_{\mathbf{s}\in\mathbb{R}^P, \|\boldsymbol{s}\|_2=1} E\left|\mathbf{s}^T\mathbf{T}_X\right|^{4\gamma+4+\varepsilon} \leq K',$$

and $\mathbf{T}_Y$ satisfies an analogous polynomial-type tail condition with sample size n_Y.

The following lemma establishes the key error bound for the empirical estimating equation when applied to our target parameter $\boldsymbol{\Delta}_0$.

Lemma 1. *Suppose $\boldsymbol{\Omega}_{X,0}, \boldsymbol{\Omega}_{Y,0} \in \mathcal{U}(M, R_X, R_Y)$ with Assumptions 1 and 2 being satisfied. Under either Assumption 3 (sub-Gaussian) or Assumption 4 (polynomial-type tail), for any $\xi > 0$,*

$$\left|\widehat{\boldsymbol{\Sigma}}_{X,c}\boldsymbol{\Delta}_0\widehat{\boldsymbol{\Sigma}}_{Y,c} - \left(\widehat{\boldsymbol{\Sigma}}_{X,c} - \widehat{\boldsymbol{\Sigma}}_{Y,c}\right)\right|_\infty \leqslant 2C\left\{M(R_X + R_Y + 1) + 1\right\}\left(\frac{\log p}{n_{\min}}\right)^{1/2} + \frac{3R_X R_Y M}{\sqrt{p}},$$

with probability at least $1 - O(p^{-\xi})$ $\left(\text{or } 1 - O(p^{-\xi/2} + n_{\min}^{-\varepsilon/8})\right)$, where $C = \max(\alpha_X, \alpha_Y)$ (or $\max(\alpha'_X, \alpha'_Y)$), as defined in Lemma A1 of the online supplementary materials, and $n_{\min} = \min\{n_X, n_Y\}$.

The bound in Lemma 1 has two components: the first term arises from sample estimation error in the compositional covariance estimators, while the second term reflects the identifiability gap between $\boldsymbol{\Delta}_0$ and $\boldsymbol{\Delta}_c$. It is important to note that this second term is

intrinsic to the compositional setting and vanishes as $p \to \infty$. Compared with the non-compositional case studied in Zhao et al. (2014), our problem explicitly incorporates this identifiability gap. Hence, Lemma 1 provides the theoretical foundation for employing a constrained ℓ_1-minimization approach in the compositional setting.

Theorem 1. *Suppose* $\mathbf{\Omega}_{X,0}, \mathbf{\Omega}_{Y,0} \in \mathcal{U}(M, R_X, R_Y)$*, with Assumptions 1 and 2 being satisfied. Under either Assumption 3 (sub-Gaussian) or Assumption 4 (polynomial-type tail), for any* $\xi > 0$*, let*

$$\lambda = 4C\left\{M(R_X + R_Y + 1) + 1\right\}\left(\frac{\log p}{n_{\min}}\right)^{1/2} + \frac{6R_X R_Y M}{\sqrt{p}}.$$

Then

$$\left|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\right|_\infty \leqslant \frac{1}{\sigma_{\min}^{S,c}}\left\{1 + \frac{\sigma_{\min}^{S,c}}{\sigma_{\min}^{S,c} - (2s-1)\mu_c}\right\} 2\lambda,$$

$$\left\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\right\|_F \leqslant \frac{(10s)^{1/2}}{\sigma_{\min}^{S,c}}\left\{1 + \frac{\sigma_{\min}^{S,c}}{\sigma_{\min}^{S,c} - (2s-1)\mu_c}\right\} 2\lambda,$$

with probability at least $1 - O(p^{-\xi})$ $\left(or\ 1 - O(p^{-\xi/2} + n_{\min}^{-\varepsilon/8})\right)$*. Here* $C = \max(\alpha_X, \alpha_Y)$ *(or* $\max(\alpha'_X, \alpha'_Y)$*), as defined in Lemma A1 of the supplementary material, and* $n_{\min} = \min\{n_X, n_Y\}$.

Theorem 1 provides finite-sample error bounds for the proposed estimator of the basis differential network matrix. The bounds are expressed under both the element-wise ℓ_∞-norm and the Frobenius norm, and they hold with high probability under either sub-Gaussian or polynomial-type tail conditions. The error bounds decompose into two terms: (i) the sample estimation error incurred when approximating the compositional covariance matrices from finite samples, and (ii) the identification error arising from the difference between $\mathbf{\Delta}_c$ and the true basis differential network $\mathbf{\Delta}_0$. Importantly, the first term, proportional to $\{\log p/\min(n_X, n_Y)\}^{1/2}$, reflects the curse of dimensionality. The second term, scaling as

$1/\sqrt{p}$, originates from the identifiability gap and captures the blessing of dimensionality, as p grows, this term vanishes. This decomposition highlights a fundamental tradeoff. On the one hand, higher dimensionality increases the sample estimation error through the $\log p$ factor; on the other hand, it simultaneously reduces the identification error via the $p^{-1/2}$ decay. As discussed in Remark 1, this $p^{-1/2}$ dimensional decay corresponds to the fixed differential-complexity regime considered in our main theory; if the differential sparsity increases with p, this dimensional benefit is partially offset by the growth of the differential-network complexity. As p becomes sufficiently large, the estimation error dominates, while the identification error becomes negligible. Thus, Theorem 1 demonstrates not only that the proposed estimator achieves consistency under mild conditions, but also that the compositional setting introduces a unique interplay between estimation and identifiability, a phenomenon absent in unconstrained differential network estimation.

Corollary 1. *Under the conditions of Theorem 1, suppose that* $n_{\min} = o(p \log p)$, *and define* $\lambda' = 4C\left\{M(R_X + R_Y + 1) + 1\right\}\left(\frac{\log p}{n_{\min}}\right)^{1/2}$. *Then the estimator* $\widehat{\mathbf{\Delta}}$ *satisfies*

$$\left|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\right|_\infty \leqslant \frac{1}{\sigma_{\min}^{S,c}}\left\{1 + \frac{\sigma_{\min}^{S,c}}{\sigma_{\min}^{S,c} - (2s-1)\mu_c}\right\} 2\lambda',$$

$$\left\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\right\|_F \leqslant \frac{(10s)^{1/2}}{\sigma_{\min}^{S,c}}\left\{1 + \frac{\sigma_{\min}^{S,c}}{\sigma_{\min}^{S,c} - (2s-1)\mu_c}\right\} 2\lambda',$$

with probability at least $1 - O(p^{-\xi})$ $\left(or\ 1 - O\left(p^{-\xi/2} + n_{\min}^{-\varepsilon/8}\right)\right)$, *where* $C = \max(\alpha_X, \alpha_Y)$ $(or\ \max(\alpha'_X, \alpha'_Y))$, *as defined in Lemma A1 of the supplementary material, and* $n_{\min} = \min\{n_X, n_Y\}$.

Corollary 1 refines Theorem 1 by imposing the additional growth condition $n_{\min} = o(p \log p)$. Under this regime, the composition-induced identification error term of order $p^{-1/2}$ becomes negligible relative to the sampling estimation error, so the error bound is

effectively dominated by $(\log p/n_{\min})^{1/2}$. As a result, the convergence rates of $\widehat{\mathbf{\Delta}}$ under both the elementwise ℓ_∞-norm and the Frobenius norm reduce to the same order as if the latent basis covariance matrices were directly observed. This reveals a form of phase transition: while Theorem 1 highlights a tradeoff between estimation and identifiability, Corollary 1 shows that in sufficiently high dimensions, the identifiability gap vanishes and the estimator achieves the same statistical efficiency as in the unconstrained (non-compositional) setting. In particular, this demonstrates that the proposed method inherits the blessing of dimensionality: as p grows, the identifiability barrier becomes asymptotically irrelevant, and the estimator performs as well as if the basis information were available.

Leveraging the ℓ_∞-error bound from Theorem 1, we define the thresholding estimator

$$\widehat{\mathbf{\Delta}}_{\tau_{n,p}} = \left\{\widehat{\delta}_{jk} \cdot I\left(|\widehat{\delta}_{jk}| > \tau_{n,p}\right)\right\}_{1\le j,k\le p}$$

for a threshold $\tau_{n,p} > 0$. Define the (j,k)th entry of $\widehat{\mathbf{\Delta}}_{\tau_{n,p}}$ as $\hat{\delta}_{jk}^{\tau_{n,p}}$, and the sign function

$$\operatorname{sgn}(t) = \begin{cases} 1 & t > 0, \\ 0 & t = 0, \\ -1 & t < 0. \end{cases}$$

Let $\mathcal{M}\big(\widehat{\mathbf{\Delta}}_{\tau_{n,p}}\big) = \left\{\operatorname{sgn}\left(\hat{\delta}_{jk}^{\tau_{n,p}}\right) : 1 \le j,k \le p\right\}$ and $\mathcal{M}(\mathbf{\Delta}_0) = \left\{\operatorname{sgn}\left(\delta_{jk}^0\right) : 1 \le j,k \le p\right\}$ represent the sign patterns of the thresholding estimator and true differential matrix, respectively. Under an additional minimum signal strength condition, the following result establishes exact sign recovery of our thresholding estimator.

Theorem 2. *Assume Theorem 1's conditions hold. If*

$$\tau_{n,p} \geqslant \frac{1}{\sigma_{\min}^{S,c}}\left\{1 + \frac{\sigma_{\min}^{S,c}}{\sigma_{\min}^{S,c} - (2s-1)\mu_c}\right\} 2\lambda,$$

and the minimum signal strength satisfies

$$\min_{j,k:\delta^0_{jk}\neq 0} \left|\delta^0_{jk}\right| > 2\tau_{n,p},$$

then $\mathcal{M}\big(\widehat{\boldsymbol{\Delta}}_{\tau_{n,p}}\big) = \mathcal{M}(\boldsymbol{\Delta}_0)$ *with probability at least* $1 - O(p^{-\xi})$ *under Assumption 3, or* $1 - O(p^{-\xi/2} + n_{\min}^{-\varepsilon/8})$ *under Assumption 4.*

Theorem 2 establishes exact support recovery of the basis differential network matrix. Specifically, it shows that if the threshold $\tau_{n,p}$ is chosen no smaller than the element-wise error bound in Theorem 1, and the true basis differential matrix $\boldsymbol{\Delta}_0$ satisfies a minimum signal strength condition ensuring that all nonzero entries are sufficiently separated from zero, then the thresholded estimator recovers the correct sign pattern with high probability. Importantly, this recovery result is based on accurate estimation and thresholding of $\boldsymbol{\Delta}_0$ and does not require the unthresholded supports of $\boldsymbol{\Delta}_0$ and $\boldsymbol{\Delta}_c$ to coincide. This result highlights that, beyond providing consistent estimation in norm, our method can reliably identify the underlying network structure, as long as the true signals are strong enough.

4. Simulation Study

We evaluate the numerical performance of DNA-Coda through comprehensive simulations and compare it against the following approaches:

- *Oracle*: Oracle version of DNA-Coda using basis vectors
- *CARE-S*: The separate-estimation approach that applies CARE to each group and then takes the difference between the two estimated precision matrices.
- *CARE-S-Oracle*: Oracle version of CARE-S using basis vectors.

Statistica Sinica: Newly accepted Paper

- *DCDTr*: The direct composition-aware differential-network estimator of He and Deng (2019), based on a lasso-penalized D-trace loss applied to the CLR-transformed observations. Following its original implementation, the tuning parameter was selected using BIC.

The proposed DNA-Coda method was implemented via the ADMM algorithm described in Section 1 of the supplementary material.

4.1 Data Generation

The compositional data $\mathbf{X}_k (k = 1, \ldots, n_X)$ and $\mathbf{Y}_{k'} (k' = 1, \ldots, n_Y)$ were simulated through the following procedure:

1. Generating log-basis vectors: $\mathbf{T}_X^{(k)} \sim N_p\left(\mathbf{0}, \mathbf{\Omega}_{X,0}^{-1}\right)$ and $\mathbf{T}_Y^{(k')} \sim N_p\left(\mathbf{0}, \mathbf{\Omega}_{Y,0}^{-1}\right)$,
2. Computing basis vectors: $\mathbf{W}_X^{(k)} = \exp\left(\mathbf{T}_X^{(k)}\right)$ and $\mathbf{W}_Y^{(k')} = \exp\left(\mathbf{T}_Y^{(k')}\right)$,
3. Calculating compositional vectors via normalization: $\mathbf{X}_k = \mathbf{W}_X^{(k)}/|\mathbf{W}_X^{(k)}|_1$ and $\mathbf{Y}_{k'} = \mathbf{W}_Y^{(k')}/|\mathbf{W}_Y^{(k')}|_1$.

We next introduce four distinct scenarios of $\mathbf{\Omega}_{X,0}$ that we have used to generate the compositional data using different network structures $\mathbf{\Delta}_0$:

Scenario 1 (Hub graph structure): Using the `huge` package, we generated a hub graph G_H with five equally-sized, non-overlapping hub modules. Nonzero off-diagonal entries in $\mathbf{\Omega}_{X,0}$ were sampled uniformly from $[-0.7, -0.5] \cup [0.5, 0.7]$. Symmetry was enforced via $\mathbf{\Omega}_{X,0} = \mathbf{Upper}_{X,0} + \mathbf{Upper}_{X,0}^{\top}$, where $\mathbf{Upper}_{X,0}$ was the upper triangular matrix of $\mathbf{\Omega}_{X,0}$.

Scenario 2 (Band graph structure): We constructed a symmetric band matrix $\mathbf{\Omega}_{X,0} = (\omega_{X,ij}^0)$ with $\omega_{X,ij}^0 = \omega_{X,ji}^0 = 0.7$ for $j = i + 1$, $\omega_{X,ij}^0 = \omega_{X,ji}^0 = 0.5$ for $j = i + 2$,

Statistica Sinica: Newly accepted Paper

$\omega^0_{X,ij} = \omega^0_{X,ji} = 0$ for $|i-j| \geq 3$, and the diagonal elements $\omega^0_{X,ii}$ were randomly selected from a uniform distribution $U(0.5, 1)$.

Scenario 3 (Block-diagonal graph structure): The p nodes were partitioned into five disjoint blocks, forming a symmetric block-diagonal precision matrix $\mathbf{\Omega}_{X,0} = (\omega^0_{X,ij})$ where for any $i \neq j$: elements between different blocks satisfied $\omega^0_{X,ij} = 0$, while within the same block, $\omega^0_{X,ij}$ was non-zero with probability $20/p$ and took value 0.5 or 0.7 with equal probability (0.5 each) when present; diagonal elements $\omega^0_{X,ii}$ were independently drawn from $U(0.5, 1)$ as specified in **Scenario 2**.

Scenario 4 (Power-law graph structure): The support of $\mathbf{\Omega}_{X,0}$ was generated from a power-law network with exponent 2 and 20% edge density ($p(p-1)/10$ edges). Nonzero off-diagonal elements $\omega^0_{X,ij}$ $(i \neq j)$ were assigned values identically to **Scenario 1**. To ensure positive definiteness, each row was scaled by 3 ($p = 40$), 4 ($p = 60$ or 80), or 5 ($p = 100$) respectively, followed by diagonal normalization ($\omega^0_{X,ii} = 1$) and symmetrization $\mathbf{\Omega}_{X,0} = \frac{1}{2}(\mathbf{\Omega}_{X,0} + \mathbf{\Omega}^\top_{X,0})$.

Differential Network Matrix Construction: For Scenarios 1-3, $\mathbf{\Omega}_{X,0}$ was made positive definite via $\mathbf{\Omega}_{X,0} = \mathbf{\Omega}_{X,0} + (|\lambda_{\min}(\mathbf{\Omega}_{X,0})| + a)\,\mathbf{I}$ $(a > 0)$, after which $\mathbf{\Omega}_{Y,0}$ was derived by inverting signs of randomly selected structural blocks (20% proportion). For Scenario 4, we inverted signs of the top 20% magnitude edges to the top two hubs to obtain $\mathbf{\Omega}_{Y,0}$. The true differential network matrix was then defined uniformly as $\mathbf{\Delta}_0 = \mathbf{\Omega}_{Y,0} - \mathbf{\Omega}_{X,0}$.

4.2 Simulation Results

To comprehensively assess estimator performance, we calculated the estimation accuracy and support recovery of each method.

Statistica Sinica: Newly accepted Paper

- *Estimation Accuracy:* Quantified through three matrix norms: Spectral Norm: $\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\|_2$, Matrix ℓ_1 Norm: $\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\|_1$, and Frobenius Norm: $\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}_0\|_{\mathrm{F}}$.

- *Support Recovery:* Evaluated for the true basis differential network matrix $\mathbf{\Delta}_0 = \left(\delta_{ij}^0\right)$ and its estimation $\widehat{\mathbf{\Delta}}_0 = \left(\widehat{\delta}_{ij}^0\right)$ via True Positive Rate (TPR) and False Positive Rate (FPR):

$$\mathrm{TPR} = \frac{\#\left\{(i,j) : \delta_{ij}^0 \neq 0 \cap \widehat{\delta}_{ij}^0 \neq 0\right\}}{\#\left\{(i,j) : \delta_{ij}^0 \neq 0\right\}}, \quad \mathrm{FPR} = \frac{\#\left\{(i,j) : \delta_{ij}^0 = 0 \cap \widehat{\delta}_{ij}^0 \neq 0\right\}}{\#\left\{(i,j) : \delta_{ij}^0 = 0\right\}}.$$

  where $\#\{\cdot\}$ denotes set cardinality, with method-specific decision thresholds applied to determine nonzero entries: our DNA-Coda method, its oracle version and DCDTr used $\tau = 0.0001$, while CARE-S and CARE-S-Oracle employed $\tau = 0.0002$.

We examined dimensions $p \in \{40, 60, 80, 100\}$ with fixed sample sizes $n_X = n_Y = 100$. Method tuning was employed by 5-fold cross-validation, with performance metrics aggregated over 100 independent replications to ensure statistical robustness. Comprehensive results for both estimation accuracy and support recovery metrics across all scenarios are presented in Tables 1–4, while ROC curves (TPR and FPR values against different λ values) are displayed in Fig. 1.

The comprehensive simulation results demonstrate the strong overall performance of DNA-Coda across all investigated scenarios and dimensionalities ($p \in \{40, 60, 80, 100\}$). Regarding estimation accuracy, DNA-Coda consistently outperforms CARE-S and CARE-S-Oracle, reducing spectral, matrix ℓ_1, and Frobenius norm losses by a margin between 35% and 70% (Tables 1-4), with particularly notable improvements in the Band structure (up to 67% spectral norm reduction) and the Block-diagonal scenario (40% matrix ℓ_1 norm improvement at $p = 100$). Compared with DCDTr, DNA-Coda generally achieves smaller

Statistica Sinica: Newly accepted Paper

Table 1: Means and standard errors (in parentheses) of performance measures for different methods in Scenario 1 (Hub graph).

p	DNA-Coda	Oracle	CARE-S	CARE-S-Oracle	DCDTr
			Spectral norm loss		
40	1.56(0.03)	1.36(0.03)	2.52(0.03)	2.47(0.03)	1.81 (0.03)
60	1.99(0.03)	1.95(0.03)	3.03(0.03)	3.20(0.03)	2.17 (0.03)
80	2.45(0.04)	2.44(0.03)	3.64(0.03)	3.81(0.03)	2.59 (0.04)
100	3.09(0.04)	3.01(0.04)	4.48(0.03)	4.50(0.03)	3.32 (0.05)
			Matrix ℓ_1-norm loss		
40	4.04(0.07)	3.46(0.07)	6.38(0.08)	6.16(0.08)	4.53 (0.07)
60	6.10(0.10)	6.02(0.10)	9.54(0.09)	10.03(0.10)	6.67 (0.12)
80	8.95(0.13)	8.95(0.14)	13.36(0.12)	14.02(0.13)	9.34 (0.16)
100	12.65(0.15)	12.38(0.15)	18.66(0.11)	18.74(0.12)	13.49 (0.25)
			Frobenius norm loss		
40	2.22(0.03)	1.95(0.03)	4.49(0.04)	4.40(0.04)	2.53 (0.04)
60	2.89(0.04)	2.83(0.04)	6.03(0.04)	5.95(0.04)	3.08 (0.05)
80	3.52(0.05)	3.51(0.05)	7.84(0.04)	7.58(0.04)	3.67 (0.06)
100	4.39(0.05)	4.27(0.05)	9.93(0.05)	9.62(0.04)	4.68 (0.08)
			True positive rate (%)		
40	99.29(0.31)	99.86(0.14)	94.14(0.98)	94.14(0.89)	97.86 (0.55)
60	97.82(0.39)	97.82(0.41)	90.82(0.97)	84.18(1.35)	95.36 (0.83)
80	95.67(0.51)	95.93(0.53)	91.00(0.75)	85.40(1.17)	93.27 (1.07)
100	92.42(0.58)	93.68(0.49)	80.37(1.00)	79.32(1.07)	86.47 (1.69)
			False positive rate (%)		
40	1.96(0.12)	2.02(0.12)	6.80(0.08)	4.65(0.06)	0.48 (0.05)
60	1.31(0.08)	1.30(0.07)	5.58(0.06)	3.78(0.04)	0.52 (0.04)
80	0.79(0.04)	0.83(0.05)	5.55(0.05)	3.76(0.03)	0.41 (0.02)
100	0.54(0.03)	0.56(0.03)	5.04(0.05)	4.02(0.04)	0.23 (0.02)

estimation losses across the spectral, matrix ℓ_1, and Frobenius norms over the investigated settings. For example, in the Block-diagonal scenario with $p = 100$, the spectral, matrix ℓ_1, and Frobenius norm losses of DNA-Coda are 3.41, 5.85, and 7.82, respectively, compared with 5.13, 8.06, and 10.54 for DCDTr. Crucially, DNA-Coda consistently approaches its oracle performance across different metrics under all graph structures, demonstrating its methodological robustness.

Statistica Sinica: Newly accepted Paper

Table 2: Means and standard errors (in parentheses) of performance measures for different methods in Scenario 2 (Band graph).

p	DNA-Coda	Oracle	CARE-S	CARE-S-Oracle	DCDTr
			Spectral norm loss		
40	1.70(0.04)	1.70(0.04)	3.23(0.02)	3.21(0.02)	2.27 (0.04)
60	2.05(0.04)	1.97(0.03)	3.34(0.02)	3.40(0.02)	2.35 (0.03)
80	2.37(0.03)	2.27(0.03)	3.35(0.02)	3.42(0.02)	2.59 (0.03)
100	2.45(0.03)	2.33(0.03)	3.50(0.02)	3.51(0.02)	3.48 (0.07)
			Matrix ℓ_1-norm loss		
40	2.81(0.06)	2.73(0.06)	4.51(0.03)	4.41(0.04)	3.01 (0.05)
60	3.16(0.06)	3.00(0.04)	4.73(0.04)	4.64(0.03)	3.15 (0.04)
80	3.64(0.06)	3.61(0.07)	4.95(0.03)	4.82(0.04)	3.48 (0.04)
100	3.65(0.06)	3.62(0.05)	5.10(0.04)	5.00(0.03)	4.24 (0.05)
			Frobenius norm loss		
40	2.85(0.04)	2.82(0.04)	5.58(0.03)	5.51(0.03)	3.51 (0.05)
60	3.88(0.04)	3.78(0.04)	6.98(0.03)	6.90(0.03)	4.19 (0.04)
80	4.93(0.04)	4.78(0.05)	8.20(0.03)	8.12(0.03)	5.25 (0.05)
100	5.31(0.04)	5.17(0.04)	9.44(0.02)	9.38(0.03)	7.77 (0.14)
			True positive rate (%)		
40	97.69(0.39)	97.69(0.43)	86.92(0.84)	85.39(0.86)	90.85 (0.76)
60	94.52(0.53)	95.86(0.43)	91.57(0.61)	88.76(0.65)	90.86 (0.53)
80	94.10(0.53)	94.52(0.50)	93.35(0.46)	90.45(0.54)	89.59 (0.61)
100	95.81(0.32)	96.08(0.31)	91.05(0.40)	90.46(0.46)	67.51 (2.49)
			False positive rate (%)		
40	2.38(0.14)	1.72(0.10)	7.22(0.13)	4.30(0.09)	0.26 (0.02)
60	1.82(0.09)	1.70(0.09)	6.67(0.09)	3.78(0.05)	0.37 (0.02)
80	1.79(0.09)	1.68(0.07)	6.67(0.07)	4.11(0.04)	0.30 (0.02)
100	0.99(0.05)	1.19(0.05)	6.24(0.06)	4.91(0.05)	0.03 (0.01)

For support recovery metrics, DNA-Coda achieves exceptional performance of maintaining high TPRs (85-99%) while constraining FPRs to $\leq 3.3\%$ across all conditions. This dual optimization substantially exceeds CARE-S and CARE-S-Oracle, which exhibit both higher FPR (4-8%) and deterioration at higher dimensions. The Power-law scenario at $p = 100$ reveals DNA-Coda's most significant advantage: a 32-36% TPR improvement (87.45-87.70% vs CARE-S and CARE-S-Oracle 51.45-54.70%) coupled with 80% FPR reduction, validated

Statistica Sinica: Newly accepted Paper

Table 3: Means and standard errors (in parentheses) of performance measures for different methods in Scenario 3 (Block-diagonal graph).

p	DNA-Coda	Oracle	CARE-S	CARE-S-Oracle	DCDTr
			Spectral norm loss		
40	1.96(0.03)	1.86(0.03)	3.79(0.03)	3.68(0.02)	2.61 (0.04)
60	2.54(0.04)	2.34(0.03)	4.04(0.03)	3.94(0.02)	3.88 (0.06)
80	2.98(0.03)	2.79(0.03)	4.23(0.03)	4.20(0.02)	4.71 (0.05)
100	3.41(0.04)	3.16(0.03)	4.61(0.02)	4.48(0.02)	5.13 (0.04)
			Matrix ℓ_1-norm loss		
40	3.38(0.06)	3.19(0.06)	6.05(0.06)	5.94(0.07)	3.73 (0.06)
60	4.18(0.07)	3.91(0.06)	6.35(0.06)	6.18(0.05)	5.38 (0.09)
80	5.01(0.07)	4.75(0.07)	6.95(0.06)	6.78(0.05)	6.46 (0.05)
100	5.85(0.07)	5.59(0.07)	8.41(0.08)	8.03(0.08)	8.06 (0.10)
			Frobenius norm loss		
40	3.46(0.04)	3.25(0.05)	6.38(0.04)	6.30(0.03)	4.06 (0.06)
60	4.96(0.05)	4.64(0.05)	8.23(0.03)	8.04(0.04)	6.76 (0.09)
80	6.52(0.05)	6.16(0.05)	10.36(0.04)	10.00(0.03)	9.15 (0.09)
100	7.82(0.05)	7.45(0.05)	12.01(0.03)	11.70(0.03)	10.54 (0.06)
			True positive rate (%)		
40	95.64(0.52)	95.86(0.52)	77.00(0.99)	74.50(0.99)	85.14 (1.19)
60	89.38(0.75)	92.05(0.68)	75.24(0.88)	73.24(0.94)	50.00 (2.49)
80	85.27(0.60)	88.30(0.51)	75.83(0.71)	72.87(0.66)	26.03 (2.31)
100	78.33(0.74)	82.44(0.59)	69.74(0.73)	71.69(0.73)	20.79 (1.85)
			False positive rate (%)		
40	3.27(0.17)	2.21(0.13)	6.38(0.11)	3.50(0.07)	0.47 (0.04)
60	2.03(0.09)	1.66(0.09)	6.01(0.07)	3.60(0.05)	0.07 (0.01)
80	1.67(0.08)	1.29(0.06)	6.61(0.07)	4.18(0.04)	0.02 (0.00)
100	1.16(0.05)	1.04(0.06)	6.23(0.05)	4.95(0.04)	0.01 (0.00)

by ROC characteristics in Fig. 1. DNA-Coda further demonstrates exceptional dimensional scalability: while CARE-S exhibits severe performance degradation (up to 39.42% TPR decline in Power-law scenarios), DNA-Coda shows systematic FPR reduction with increasing dimensionality (e.g., 1.96% $\rightarrow$ 0.54% in Hub structure). Scenario-specific analyses reveal complementary strengths, including near-perfect Hub recovery (99.29% TPR at $p = 40$), stable Band structure performance, and maximum Frobenius norm improvements in the

Statistica Sinica: Newly accepted Paper

Table 4: Means and standard errors (in parentheses) of performance measures for different methods in Scenario 4 (Power-law graph).

p	DNA-Coda	Oracle	CARE-S	CARE-S-Oracle	DCDTr
			Spectral norm loss		
40	0.46(0.01)	0.44(0.01)	0.87(0.01)	0.95(0.01)	0.59 (0.01)
60	0.55(0.01)	0.53(0.01)	0.96(0.01)	1.01(0.01)	0.64 (0.01)
80	0.64(0.01)	0.63(0.01)	1.13(0.01)	1.14(0.01)	0.72 (0.01)
100	0.73(0.01)	0.73(0.01)	1.13(0.01)	1.13(0.01)	0.77 (0.01)
			Matrix ℓ_1-norm loss		
40	1.20(0.02)	1.20(0.03)	2.71(0.03)	2.91(0.03)	1.46 (0.03)
60	1.86(0.03)	1.81(0.03)	3.62(0.03)	3.75(0.03)	2.01 (0.03)
80	2.51(0.04)	2.49(0.04)	4.62(0.03)	4.65(0.03)	2.61 (0.04)
100	3.11(0.03)	3.11(0.03)	5.10(0.02)	5.10(0.02)	3.26 (0.04)
			Frobenius norm loss		
40	0.63(0.01)	0.61(0.01)	1.62(0.01)	1.70(0.01)	0.85 (0.01)
60	0.78(0.01)	0.76(0.01)	2.00(0.02)	1.95(0.01)	1.05 (0.01)
80	0.90(0.01)	0.89(0.01)	2.51(0.02)	2.33(0.01)	1.04 (0.01)
100	1.03(0.01)	1.02(0.01)	2.81(0.01)	2.69(0.01)	1.17 (0.01)
			True positive rate (%)		
40	87.50(0.00)	87.73(0.13)	85.12(0.58)	81.63(0.82)	84.75 (0.60)
60	90.08(0.35)	90.17(0.36)	78.42(1.05)	69.17(1.23)	85.08 (0.98)
80	91.69(0.35)	92.00(0.35)	74.50(1.11)	69.75(1.29)	86.06 (0.67)
100	87.45(0.56)	87.70(0.62)	54.70(1.18)	51.45(1.22)	78.45 (0.99)
			False positive rate (%)		
40	1.62(0.09)	1.73(0.08)	3.82(0.08)	2.34(0.05)	0.90 (0.07)
60	1.09(0.04)	1.25(0.05)	3.40(0.06)	2.03(0.04)	1.18 (0.06)
80	0.91(0.03)	0.93(0.03)	3.92(0.05)	2.23(0.03)	0.42 (0.03)
100	0.64(0.02)	0.67(0.02)	4.14(0.04)	3.11(0.04)	0.64 (0.04)

Block-diagonal structure. Compared with DCDTr, DNA-Coda achieves a higher TPR in every investigated setting, whereas DCDTr often yields a lower FPR, indicating a more conservative edge-selection pattern. The difference is particularly pronounced in the Block-diagonal scenario with $p = 100$, where the TPRs of DNA-Coda and DCDTr are 78.33% and 20.79%, respectively, and in the Band scenario with $p = 100$, where the corresponding TPRs are 95.81% and 67.51%. Under the Power-law structure with $p = 100$, DNA-Coda

also achieves a higher TPR than DCDTr (87.45% versus 78.45%), while both methods have the same FPR of 0.64%. The ROC curves in Fig. 1 further support these findings, showing that DNA-Coda generally provides a more favorable overall trade-off between TPR and FPR than DCDTr across the investigated graph structures and dimensionalities.

Collectively, these results position DNA-Coda as an efficient statistical method for differential network analysis of compositional data, delivering: (1) superior estimation accuracy across different data settings; (2) favorable support recovery balancing sensitivity and specificity; and (3) exceptional dimensional scalability. These attributes prove to be particularly valuable for high-dimensional biological applications.

5. Application to mucosal microbiome study

To demonstrate the potential usefulness of our proposed method, we analyzed a mucosal microbiome dataset collected from 255 inflammatory bowel disease patients (Morgan et al., 2015). In this mucosal microbiome study, 16S rRNA gene sequencing was employed to measure the microbial community. These raw sequence counts were first clustered into operational taxonomic units (OTUs) and then aggregated into the genus level. Besides microbiome compositions, information of some clinical variables such as disease subtype was also collected. In particular, these 255 inflammatory bowel disease patients were divided into two groups: familial adenomatous polyposis (FAP group, $n_1 = 39$) and non-FAP group ($n_2 = 216$). It is often of interest to study the microbial conditional dependence relationships between different clinical conditions. In particular, we are interested in conducting the proposed differential network analysis between two disease subtypes. To further mitigate potential impact of sequencing measurement errors, we eliminated taxa that were overly

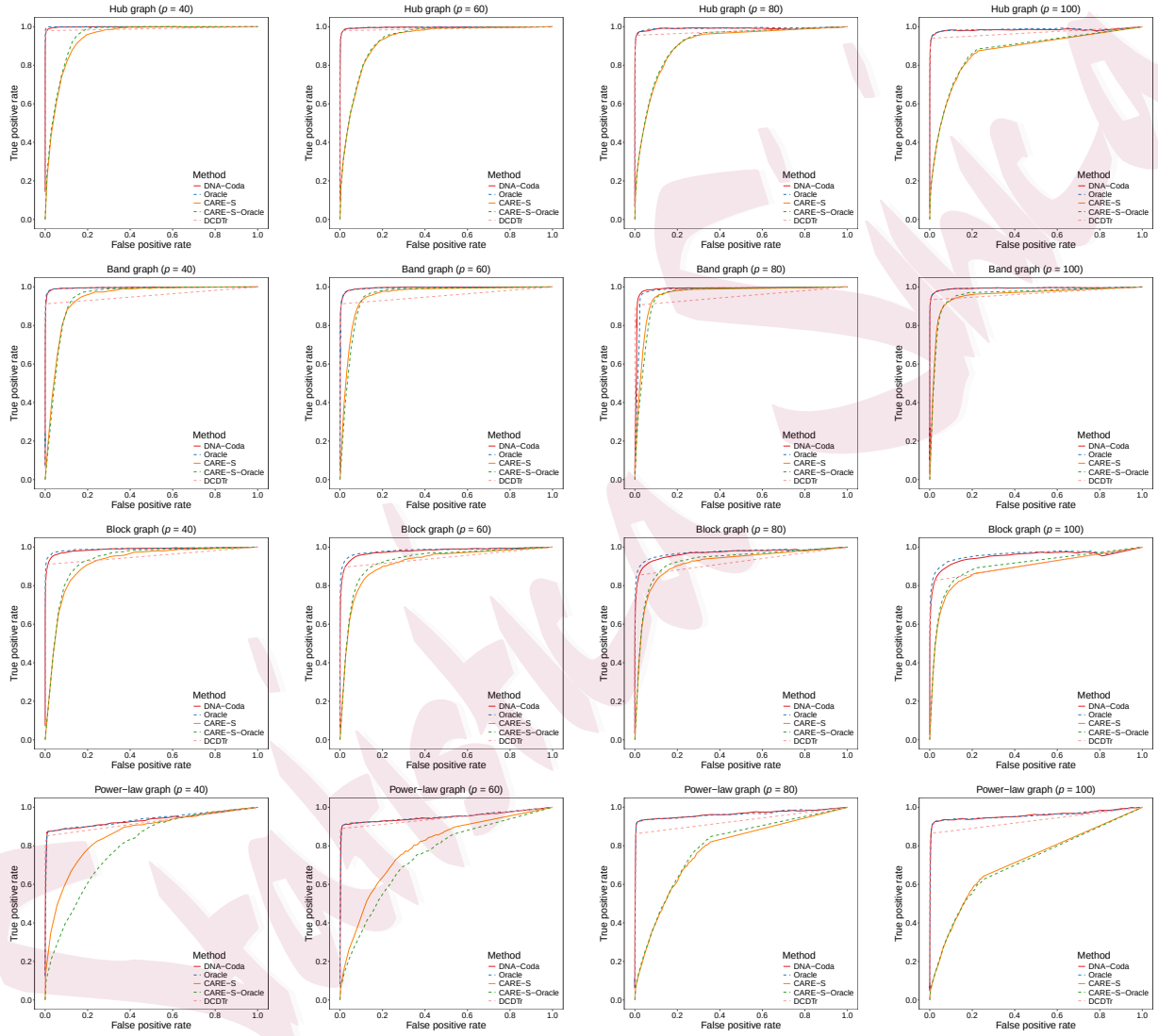


Figure 1: The ROC curves for different methods in models (a)–(d) with p = 40, 60, 80, 100.

sparse and kept the remaining $p = 62$ taxa which were present in at least 10% of the samples within both the FAP group and the non-FAP group. Finally, raw sequence counts data were transformed into compositions after adding 0.5 to all counts as suggested by a previous investigation (Shi et al., 2022).

We compared DNA-Coda with CARE-S and DCDTr using the tuning procedures described in Section 4. For numerical edge selection, thresholds of (1×10^{-4}) were applied to DNA-Coda and DCDTr, whereas a threshold of (2×10^{-4}) was used for CARE-S. The results are presented in the form of microbial interaction differential networks among the genera. To evaluate the stability of support recovery, we applied bootstrap resampling and noted the proportion of edges that reappeared in each bootstrap sample. We repeated this bootstrap process 100 times and used the average proportion of the reproduced edges as an indicator of differential network stability. Eventually, only those edges that appeared in at least 90% of the bootstrap samples were retained in the networks as stable edges.

The differential networks identified by DNA-Coda and CARE-S are shown in Figure 2. At the bootstrap selection-frequency threshold of 0.90, DCDTr identifies no stable differential edge; therefore, no DCDTr network is displayed in Figure 2. Based on this figure, the differential network detected by DNA-Coda is sparser (14 stable edges) than that of CARE-S (42 stable edges). Besides sizes of networks, the differential network stability scores of DNA-Coda, CARE-S, and DCDTr are 0.7, 0.578, and 0.533, respectively. That is, DNA-Coda tends to produce a more stable differential network than both CARE-S and DCDTr. However, this high bootstrap edge-wide reproducibility of DNA-Coda might also be contributed from the fact that it select fewer edges that the other two methods.

We next take a closer look at the differential network between FAP and non-FAP by

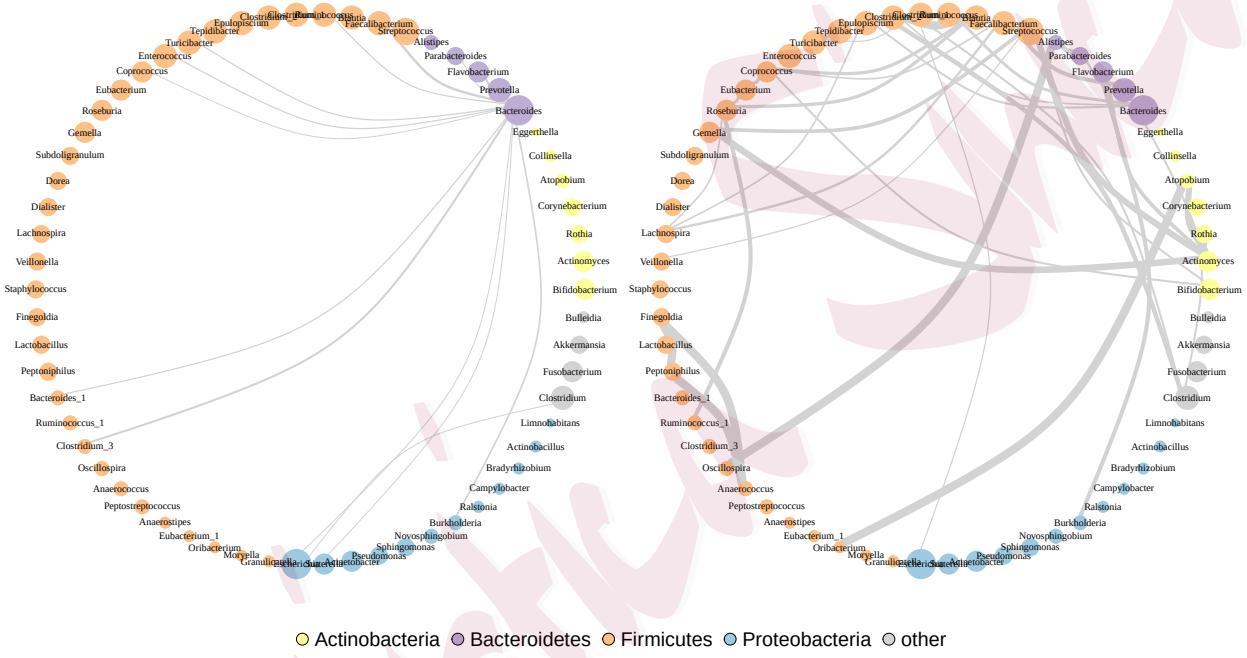


Figure 2: Differential networks between major genera of FAP group and non-FAP group identified by DNA-Coda (Left) and CARE-S (Right). Thicknesses of edges are proportional to their strengths and node sizes are proportional to the relative abundances of genera.

DNA-Coda as displayed in the left panel of Figure 2. Bacteroides and Escherichia are the only two connected hub nodes that are involved in more than one edge which are different between two conditions, and both Bacteroides and Escherichia were identified as playing pivotal roles in previous differential analyses (Morgan et al., 2015; Li et al., 2023). These observations provide useful biological context but do not independently validate the specific differential network edges identified here, and further studies using independent cohorts and complementary experimental evidence are needed to assess the biological relevance of these differential associations. On the other hand, there are more than ten hub nodes that have more than two connections in the differential network presented in the right panel of Figure 2, and no clear pattern was identified by CARE-S.

6. Discussion

High-dimensional compositional data are frequently encountered in many fields, especially in many next-generation sequencing-based studies (e.g., 16S rRNA microbiome sequencing experiments and single-cell RNA sequencing studies) that share a similar measurement process in which the total abundance information is lost and the sequence count data reflect only the relative abundances (i.e., compositional) information of unique sequences of interest (Vandeputte et al., 2017). Statistical learning on the component-component interaction network from these data is challenging both due to the closure effects of components within a compositional vector and its high dimensionality. We propose a new approach for differential network analysis of high-dimensional compositional data. Instead of separately estimating two basis precision matrices, our proposed method directly estimates their differences. Theoretical guarantees of the proposed method are provided for the basis differential network

Statistica Sinica: Newly accepted Paper

matrix estimation and support recovery.

Our approach relies on the identifiability results of the latent basis precision matrix using the observed compositional precision matrix established in the CARE method (Zhang et al., 2025). Extensions to the differential network analysis framework of DNA-Coda and also to establish theoretical guarantees of this new approach are non-trivial but rewarding. As it has been clearly demonstrated in our numerical studies that the proposed DNA-Coda method substantially outperforms the competing approach that applies CARE twice for each compositional matrix both in terms of network estimation and support recovery. The proposed method relies on the assumption that differences of two precision matrices are sparse to maintain desired properties and good performance in high dimensions. In presence of clear evidence on departure from this assumption, further investigation on alternative approaches may be desired.

Supplementary Material

The Supplementary Material contains implementation details of DNA-Coda and technical proofs of theorems.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological) 44*(2), 139–160.

Aitchison, J. (2003). The statistical analysis of compositional data. *Caldwell: The Blackburn Press*.

Bandyopadhyay, S., M. Mehta, D. Kuo, M.-K. Sung, R. Chuang, E. J. Jaehnig, B. Bodenmiller, K. Licon, W. Copeland, M. Shales, et al. (2010). Rewiring of genetic networks in response to dna damage. *Sci-*

Statistica Sinica: Newly accepted Paper

ence 330(6009), 1385–1389.

Brill, B., A. Amir, and R. Heller (2022). Testing for differential abundance in compositional counts data, with application to microbiome studies. *Annals of Applied Statistics 16*(4), 2648–2671.

Cai, T., H. Li, J. Ma, and Y. Xia (2019). Differential markov random field analysis with an application to detecting differential microbial community networks. *Biometrika 106*(2), 401–416.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 594–607.

Cai, T. T., W. Liu, and H. H. Zhou (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics 44*, 455–488.

Cao, Y., W. Lin, and H. Li (2018). Two-sample tests of high-dimensional means for compositional data. *Biometrika 105*(1), 115–132.

Cao, Y., W. Lin, and H. Li (2019). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association 114*(526), 759–772.

Danaher, P., P. Wang, and D. M. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(2), 373–397.

Fang, H. (2023). gmcoda: Graphical model for multiple compositional vectors in microbiome studies. *Bioinformatics 39*(11), btad700.

Fang, H., C. Huang, H. Zhao, and M. Deng (2017). gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology 24*(7), 699–708.

Fernandes, A. D., J. N. Reid, J. M. Macklaim, et al. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome 2*, 1–13.

He, S. and M. Deng (2019). Direct interaction network and differential network inference from compositional data

Statistica Sinica: Newly accepted Paper

via lasso penalized d-trace loss. *Plos one 14*(7), e0207731.

Jiang, L., X. Liu, X. He, et al. (2021). A behavioral model for mapping the genetic architecture of gut-microbiota networks. *Gut Microbes 13*(1), 1820847.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, et al. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology 11*(5), e1004226.

Li, D., A. Srinivasan, L. Xue, and X. Zhan (2023). Robust shape matrix estimation for high-dimensional compositional data with application to microbial inter-taxa analysis. *Statistica Sinica 33*, 1577–1602.

Morgan, X. C., B. Kabakchiev, L. Waldron, et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome biology 16*, 1–15.

Shi, P., Y. Zhou, and A. R. Zhang (2022). High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. *Biometrika 109*(2), 405–420.

Vandeputte, D., G. Kathagen, K. D'hoe, et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature 551*(7681), 507–511.

Yuan, H., R. Xi, C. Chen, and M. Deng (2017). Differential network analysis via lasso penalized d-trace loss. *Biometrika 104*(4), 755–770.

Zhang, S., H. Wang, and W. Lin (2025). Care: Large precision matrix estimation for compositional data. *Journal of the American Statistical Association 120*(549), 305–317.

Zhao, S. D., T. T. Cai, and H. Li (2014). Direct estimation of differential networks. *Biometrika 101*(2), 253–268.

Zhou, H., K. He, J. Chen, and X. Zhang (2022). Linda: linear models for differential abundance analysis of microbiome compositional data. *Genome biology 23*(1), 95.

Statistica Sinica: Newly accepted Paper

Data Science Institute, Shandong University

E-mail: machencheng@mail.sdu.edu.cn

Hao chen

School of Public Health, Cheeloo College of Medicine, Shandong University

E-mail: hao.chen@sdu.edu.cn

Yijuan Hu

Beijing International Center for Mathematical Research, Peking University

E-mail: yijuanhu@bicmr.pku.edu.cn

Xiang Zhan

School of Statistics and Data Science, Southeast University

E-mail: zhanx@seu.edu.cn

Statistica Sinica