Statistica Sinica Preprint No: SS-2026-0002	
Title	Precision Matrix Estimation for Multiple Compositional Vectors
Manuscript ID	SS-2026-0002
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0002
Complete List of Authors	Shen Zhang, Huaying Fang and Tao Hu
Corresponding Authors	Huaying Fang
E-mails	hyfang@cnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Precision Matrix Estimation for Multiple Compositional Vectors

Shen Zhang[1], Huaying Fang[2,] and Tao Hu[2]

[1] *Qufu Normal University and* [2] *Capital Normal University*

Abstract: Compositional data represent the relative abundances of individual components within a whole and arise in a wide range of scientific fields. A central task in compositional data analysis is to characterize conditional dependence among components, typically through estimation of the precision matrix in graphical models. Most existing methods are designed for a single compositional vector. In microbiome studies, however, multiple compositional vectors — such as bacterial and fungal profiles — are often observed for each sample, posing additional challenges for network inference. In particular, precision matrix estimation is complicated by compositional constraints, high dimensionality and the need to jointly model multiple interdependent compositional vectors. In this paper, we propose an adaptive procedure, called Amcoda, for estimating the precision matrix of latent absolute abundances from multiple compositional vectors. We establish convergence rates for the Amcoda estimator under various matrix norms and provide theoretical guarantees for support recovery. Extensive simulations demonstrate that Amcoda outperforms existing methods in both estimation accuracy and edge detection for network inference from multiple

compositional vectors. We further apply Amcoda to a real microbiome dataset to infer bacterial-fungal cross-domain networks and identify several interactions of potential biological interest.

Key words and phrases: Compositional data analysis, graphical models, high-dimensional inference, microbiome networks, precision matrix estimation.

1. Introduction

Compositional data lie in the positive simplex and represent relative proportions of components, such as microbial abundances in biology, sector contributions in economics and word frequencies in text mining. Understanding conditional dependence — typically characterized by the inverse covariance matrix (also called precision matrix) — is crucial for uncovering complex interactions among components, for example, microbe-microbe relationships that may influence host health in microbiome studies. While most existing methods focus on a single compositional vector, many modern applications generate multiple compositional vectors per sample. This setting brings several challenges: (i) the constant-sum constraint, which invalidates classical correlation analyses and can induce spurious dependencies; (ii) high dimensionality, which exacerbates both statistical and computational difficulties in precision matrix estimation; and (iii) the need

to jointly model interdependent compositional vectors (e.g., bacterial and fungal abundances) while respecting multiple simplex constraints.

Several methodological directions have been developed for a single compositional vector. Early approaches rely on approximating the covariance structure via the centered log-ratio (clr) transformation. SPIEC-EASI combines the clr transformation with neighborhood selection or graphical lasso to infer the precision matrix (Kurtz et al. , 2015). Its key assumption — that the clr covariance approximates the covariance of latent absolute abundances — may fail for certain network structures. Related clr-based covariance approximations have also been combined with SCIO-type estimators (Liu and Luo , 2015; Zhang and He , 2019). A second line of work develops likelihood-based methods. GCoda imposes a logistic-normal model on compositional data and estimates the precision matrix via penalized maximum likelihood (Fang et al. , 2017); to address the resulting non-convex optimization, gCoda employs a majorization-minimization algorithm to improve computational efficiency and stability relative to SPIEC-EASI. More recent contributions introduce regularized and robust approaches: one replaces the sample covariance in ACLIME with a Huber-type estimator (Liang, Wu and Ma , 2022; Cai, Liu and Luo , 2011), and CARE establishes asymptotic identifiability of the precision ma-

trix of latent absolute abundances and adapts CLIME to the compositional setting with convergence guarantees and provable support recovery (Zhang, Wang and Lin , 2025).

Most existing methods, however, are designed for a single compositional vector. In practice, multiple compositional vectors routinely arise. For example, in microbiome research, bacterial abundances are typically quantified using 16S rRNA sequencing, whereas fungal abundances are measured using internal transcribed spacer (ITS) sequencing, yielding two compositional vectors for each sample. Extending single-vector approaches to the multi-vector setting poses at least two challenges. First, the constraint structure changes fundamentally: the single-vector case imposes one simplex constraint (components sum to 1), whereas the multi-vector case imposes multiple simplex constraints — one for each compositional vector — so multiple constraints must be handled simultaneously. Second, both theory and computation become more challenging. Correlation across compositional vectors complicates the derivation of key estimator properties relative to the single-vector setting, and the computational burden increases due to higher dimensionality and the need to account for multiple constraints. Tipton et al. (2018) extends SPIEC-EASI to infer bacterial-fungal interactions, but this approach inherits the limitations of clr-based approximations

and may fail under certain conditions. More recently, gmcoda uses an additive logistic-normal model to estimate partial correlations across domains and performs well empirically (Fang , 2023). Nevertheless, general statistical methods for precision matrix inference with multiple compositional vectors remain underdeveloped.

In this paper, we propose an adaptive procedure, called Amcoda, for estimating the precision matrix of latent absolute abundances from multiple compositional vectors. Amcoda extends single-vector methodology to the multi-vector setting and provides theoretical guarantees including convergence rates and support recovery. Through comprehensive simulations, we show that Amcoda outperforms competing approaches in both estimation accuracy and edge detection, particularly in high-dimensional regimes. We further apply Amcoda to microbiome data to infer bacterial-fungal networks and identify several interactions of potential biological interest.

This manuscript is organized as follows. Section 2 introduces the model specification and presents the proposed methodology. Section 3 establishes the theoretical properties of Amcoda including convergence rates and support recovery. Section 4 reports simulation studies and an application to microbial cross-domain network inference. Section 5 concludes with a discussion. Proofs of all theoretical results and additional simulation results

are provided in the Supplementary Material.

2. Methods

2.1 Preliminaries

We begin by introducing the notations and definitions used throughout. All vectors are column vectors. For a matrix $\mathbf{A}$, let $\mathbf{A}^\top$ denote its transpose. Let $\mathbf{I}_p$, $\mathbf{1}_p$ and $\boldsymbol{e}_j$ for $j = 1, \ldots, p$ denote the $p \times p$ identity matrix, the p-vector of ones and the p-vector with a 1 in the jth position and 0s elsewhere, respectively. For $1 \leq \alpha \leq \infty$, let $\|\boldsymbol{a}\|_\alpha$ denotes the vector ℓ_α-norm for a vector $\boldsymbol{a}$. For a matrix $\mathbf{A}$, let $\|\mathbf{A}\|_\alpha$, $\|\mathbf{A}\|_{\max}$ and $\|\mathbf{A}\|_{\mathrm{F}}$ denote the matrix ℓ_α-norm, the entrywise max-norm and the Frobenius norm, respectively. For a vector $\boldsymbol{a} = (a_i)$ and a matrix $\mathbf{A} = (a_{ij})$, define $\exp(\boldsymbol{a}) = (\exp(a_i))$, $\log(\boldsymbol{a}) = (\log(a_i))$ and $\mathrm{sgn}(\mathbf{A}) = (\mathrm{sgn}(a_{ij}))$ where $\mathrm{sgn}(\cdot)$ denotes the sign function. For a positive semidefinite matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$ and $\lambda^+_{\min}(\mathbf{A})$ denote, respectively, (i) the smallest eigenvalue, (ii) the largest eigenvalue and (iii) the smallest strictly positive eigenvalue of $\mathbf{A}$.

Assume that d compositional vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}$ are observed where

$$\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_{p_i}^{(i)})^\top, \quad \sum_{j=1}^{p_i} X_j^{(i)} = 1, \quad i = 1, \ldots, d.$$

Let $p = \sum_{i=1}^d p_i$ and without loss of generality assume $p_1 \geq \ldots \geq p_d$. Each

compositional vector $\boldsymbol{X}^{(i)}$ is linked to an unobserved log-basis vector $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, \ldots, Y_{p_i}^{(i)})^\top$ through

$$X_j^{(i)} = \frac{\exp(Y_j^{(i)})}{\mathbf{1}_{p_i}^\top \exp(\boldsymbol{Y}^{(i)})}, \quad j = 1, \ldots, p_i, \ i = 1, \ldots, d. \tag{2.1}$$

Define the concatenated vectors

$$\boldsymbol{X} = \left((\boldsymbol{X}^{(1)})^\top, \ldots, (\boldsymbol{X}^{(d)})^\top\right)^\top, \ \boldsymbol{Y} = \left((\boldsymbol{Y}^{(1)})^\top, \ldots, (\boldsymbol{Y}^{(d)})^\top\right)^\top.$$

Our goal is to infer the conditional dependence structure of the latent vector $\boldsymbol{Y}$ based on observations of $\boldsymbol{X}$. Let $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0^{-1} = (\omega_{kl}^0)$ denote the covariance matrix and the precision matrix of $\boldsymbol{Y}$, respectively. If $\boldsymbol{Y}$ follows a multivariate normal distribution, then the conditional independence between the kth and lth variables of $\boldsymbol{Y}$ (given the other $p-2$ variables) is equivalent to $\omega_{kl}^0 = 0$. Thus, $\boldsymbol{\Omega}_0$ encodes the conditional dependence network under multivariate normality of $\boldsymbol{Y}$.

Taking logarithms on both sides of Equation (2.1) gives

$$\log(X_j^{(i)}) = Y_j^{(i)} - \log\left\{\mathbf{1}_{p_i}^\top \exp\left(\boldsymbol{Y}^{(i)}\right)\right\}, \quad j = 1, \ldots, p_i, \ i = 1, \ldots, d.$$

Let $\mathbf{R} = \mathrm{diag}(\mathbf{1}_{p_1}, \ldots, \mathbf{1}_{p_d})$ and define the block-diagonal projection matrix $\mathbf{G} = \mathbf{I}_p - \mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top$. Then

$$\mathbf{G}\log(\boldsymbol{X}) = \mathbf{G}\boldsymbol{Y}. \tag{2.2}$$

Let $\boldsymbol{\Sigma}_c = (\sigma^c_{kl})$ denote the covariance matrix of $\mathbf{G}\log(\boldsymbol{X})$. From Equation (2.2),

$$\boldsymbol{\Sigma}_c = \mathbf{G}\boldsymbol{\Sigma}_0\mathbf{G}. \tag{2.3}$$

Moreover, for $d \geq 1$,

$$\|\boldsymbol{\Sigma}_c - \boldsymbol{\Sigma}_0\|_{\max} \leq 3\|\boldsymbol{\Sigma}_0\|_1/p_d.$$

In particular, if $\|\boldsymbol{\Sigma}_0\|_1 = o(p_d)$, then $\|\boldsymbol{\Sigma}_c - \boldsymbol{\Sigma}_0\|_{\max} = o(1)$. This suggests approximating $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0^{-1}$ by an "inverse" of $\boldsymbol{\Sigma}_c$. However, due to the d constant-sum constraints, $\boldsymbol{\Sigma}_c$ is singular. We therefore work with its Moore-Penrose inverse $\boldsymbol{\Omega}_c = (\omega^c_{kl})$. The relationship among $\boldsymbol{\Sigma}_c$, $\boldsymbol{\Omega}_c$ and the target $\boldsymbol{\Sigma}_0$ is summarized in Proposition 1 (proved in Supplementary Text S1).

Proposition 1. *Let $\boldsymbol{\Omega}_c$ be the Moore-Penrose inverse of $\boldsymbol{\Sigma}_c$. Then*

$$\boldsymbol{\Omega}_c = \boldsymbol{\Omega}_0 - \boldsymbol{\Omega}_0\mathbf{R}\left(\mathbf{R}^\top\boldsymbol{\Omega}_0\mathbf{R}\right)^{-1}\mathbf{R}^\top\boldsymbol{\Omega}_0,$$

and

(a) $\boldsymbol{\Sigma}_c\boldsymbol{\Omega}_c = \mathbf{G}$;

(b) $\mathbf{G}\boldsymbol{\Omega}_c = \boldsymbol{\Omega}_c$ *and* $\boldsymbol{\Omega}_c\mathbf{R} = \mathbf{0}$;

(c) $\lambda_{\min}(\boldsymbol{\Omega}_0) \leq \lambda^+_{\min}(\boldsymbol{\Omega}_c) \leq \lambda_{\max}(\boldsymbol{\Omega}_c) \leq \lambda_{\max}(\boldsymbol{\Omega}_0)$;

(d) $\min_{1\le k\le p}\sigma^c_{kk}\omega^c_{kk} \ge \left(1-1/p_d\right)^2$.

Proposition 1 shows that $\mathbf{\Omega}_c$ behaves similarly to $\mathbf{\Omega}_0$, up to a rank-d correction induced by the simplex constraints. In particular, part (a) implies that $\mathbf{\Sigma}_c\mathbf{\Omega}_c$ equals $\mathbf{G}$ rather than the identity matrix. Part (b) follows from the special structure of $\mathbf{G}$ and the identity $\mathbf{R}^\top\mathbf{G}=\mathbf{0}$. Part (c) relates the positive eigenvalues of $\mathbf{\Omega}_c$ to those of $\mathbf{\Omega}_0$, and part (d) provides a diagonal-dominance-type bound linking $\mathbf{\Sigma}_c$ and $\mathbf{\Omega}_c$.

To study identifiability and convergence, define the matrix class

$$\mathcal{U}_{q,r}\left(s_0(p),M_p\right)=\{\mathbf{\Omega}=(\omega_{kl}):\mathbf{\Omega}\succ 0,\ \max_{1\le l\le p}\sum_{k=1}^{p}|\omega_{kl}|^q\le s_0(p),\|\mathbf{\Omega}\|_1\le M_p,$$
$$1/r\le\lambda_{\min}(\mathbf{\Omega})\le\lambda_{\max}(\mathbf{\Omega})\le r\},$$

where $\mathbf{\Omega}\succ 0$ denotes positive definiteness, $0\le q<1$ and $r>1$ are constants, and $s_0(p)$ and M_p may diverge with p. If $\mathbf{\Omega}_0\in\mathcal{U}_{q,r}\left(s_0(p),M_p\right)$, then the following properties hold (proved in Supplementary Text S2).

Proposition 2. *Suppose* $\mathbf{\Omega}_0\in\mathcal{U}_{q,r}\left(s_0(p),M_p\right)$. *Then*

(a) $\frac{p_d}{rp_1^2}\le\|\mathbf{\Omega}_c-\mathbf{\Omega}_0\|_{\max}\le\frac{rM_p^2}{p_d}$;

(b) $\|\mathbf{\Omega}_c\|_1\le\left(1+\frac{r^2\sqrt{d}p_1}{p_d}\right)M_p$;

(c) $1/r\le\lambda^+_{\min}\left(\mathbf{\Omega}_c\right)\le\lambda_{\max}\left(\mathbf{\Omega}_c\right)\le r$.

Proposition 2(a) quantifies the identification error incurred when using $\boldsymbol{\Omega}_c$ as a surrogate for $\boldsymbol{\Omega}_0$. In particular, if $M_p = o(\sqrt{p_d})$, then this error vanishes as $p_d \to \infty$, so that $\boldsymbol{\Omega}_c$ provides an asymptotically accurate approximation to $\boldsymbol{\Omega}_0$. Parts (b) and (c) show that the ℓ_1-norm and positive eigenvalues of $\boldsymbol{\Omega}_c$ remain bounded within the same matrix class $\mathcal{U}_{q,r}\left(s_0(p), M_p\right)$.

2.2 Amcoda: Adaptive estimation of precision matrix from multiple compositional vectors

Suppose we observe independent and identically distributed samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $\boldsymbol{X}$ from the representation in Equation (2.1). The sample covariance estimator of $\boldsymbol{\Sigma}_c$ is

$$\widehat{\boldsymbol{\Sigma}}_c = \frac{1}{n-1}\sum_{s=1}^{n} \mathbf{G}\left\{\log(\boldsymbol{X}_s) - \frac{1}{n}\sum_{l=1}^{n}\log(\boldsymbol{X}_l)\right\}\left\{\log(\boldsymbol{X}_s) - \frac{1}{n}\sum_{l=1}^{n}\log(\boldsymbol{X}_l)\right\}^{\top}\mathbf{G}.$$

By Equation (2.3), $\widehat{\boldsymbol{\Sigma}}_c$ consistently estimates $\mathbf{G}\boldsymbol{\Sigma}_0\mathbf{G}$. Propositions 1-2 further imply that both $\widehat{\boldsymbol{\Sigma}}_c\boldsymbol{\Omega}_c$ and $\widehat{\boldsymbol{\Sigma}}_c\boldsymbol{\Omega}_0$ concentrate around $\mathbf{G}$. This motivates estimating $\boldsymbol{\Omega}_0$ via a constrained ℓ_1-minimization procedure:

$$\widetilde{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}} \|\boldsymbol{\Omega}\|_1 \quad \text{s.t.} \quad \|\widehat{\boldsymbol{\Sigma}}_c\boldsymbol{\Omega} - \mathbf{G}\|_{\max} \le \lambda, \tag{2.4}$$

where $\lambda > 0$ is a tuning parameter. If compositional constraints are ignored, the procedure (2.4) reduces to the classical CLIME estimator. However, the

solution $\widetilde{\mathbf{\Omega}} = (\tilde{\omega}_{kl})$ is not guaranteed to be symmetric. We therefore define the symmetrized Amcoda estimator $\widehat{\mathbf{\Omega}} = (\widehat{\omega}_{kl})$ by

$$\widehat{\omega}_{kl} = \widehat{\omega}_{lk} = \widetilde{\omega}_{kl} I\left(|\widetilde{\omega}_{kl}| \leq |\widetilde{\omega}_{lk}|\right) + \widetilde{\omega}_{lk} I\left(|\widetilde{\omega}_{kl}| > |\widetilde{\omega}_{lk}|\right),$$

where $I(\cdot)$ denotes the indicator function. The resulting matrix $\widehat{\mathbf{\Omega}}$ is our estimator of $\mathbf{\Omega}_0$.

For numerical computation and tuning parameter selection, Amcoda adopts the optimization framework and tuning strategy developed in CARE, which ensures computational stability and aligns with our theoretical analysis (Supplementary Text S7).

3. Theoretical properties

In this section, we establish theoretical guarantees for the Amcoda estimator, focusing on its convergence rates and support recovery under suitable conditions. Without loss of generality, we assume that each component of the log-basis vector $\boldsymbol{Y}$ has mean 0. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ denote the column vectors of $\mathbf{G}\mathbf{\Omega}_0$. To derive finite-sample guarantees for estimating the precision matrix $\mathbf{\Omega}_0$, we impose tail conditions on $\boldsymbol{Y}$.

Condition 1.

(i) (*Sub-Gaussian tails*) There exist constants $\eta > 0$ and $K \geq 1$ such

that

$$\max_{\boldsymbol{a}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_p,\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}} E\left[\exp\left\{\eta\left(\frac{\boldsymbol{a}^\top\boldsymbol{Y}}{\|\boldsymbol{a}\|_2}\right)^2\right\}\right] \le K.$$

(ii) (*Polynomial-type tails*) There exist constants $\tau > 4$ and $K \ge 0$ such that

$$\max_{\boldsymbol{a}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_p,\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}} E\left(\left|\frac{\boldsymbol{a}^\top\boldsymbol{Y}}{\|\boldsymbol{a}\|_2}\right|^\tau\right) \le K.$$

Compared with the tail conditions assumed for a single compositional vector in Zhang, Wang and Lin (2025), Condition 1 is weaker. Specifically, Zhang, Wang and Lin (2025) requires the corresponding bounds to hold for *all* unit vectors in $\mathbb{R}^p$, whereas Condition 1 restricts the requirement to the particular set of vectors by the columns of $\mathbf{I}_p$ and $\mathbf{G}\boldsymbol{\Omega}_0$. The next theorem provides the convergence rate of the Amcoda estimator.

Theorem 1. *Suppose* $\boldsymbol{\Omega}_0 \in \mathcal{U}_{q,r}\left(s_0(p), M_p\right)$ *with* $M_p = o\left(\sqrt{p}\right)$.

(a) If $\boldsymbol{Y}$ *satisfies Condition 1(i), then for any* $\theta > 0$, *there exist constants* $C_1 > 0$, $\eta_1 > 0$ *and* $\eta_2 > 0$ *such that, if* $\frac{\log p}{n} \le \eta_2$ *and*

$$\lambda = \frac{\sqrt{p}+1}{p_d \lambda_{\min}\left(\boldsymbol{\Omega}_0\right)}\|\boldsymbol{\Omega}_0\|_1 + \eta_1\sqrt{\frac{\log p}{n}},$$

then the Amcoda estimator $\widehat{\boldsymbol{\Omega}}$ *satisfies*

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\max} \le C_1 M_p\left(\sqrt{\frac{\log p}{n}} + M_p\frac{\sqrt{p}}{p_d}\right) \tag{3.1}$$

with probability at least $1 - O\left(p^{-\theta}\right)$.

(b) *If* $\boldsymbol{Y}$ *satisfies Condition 1(ii) and* $p \leq Cn^{\tau_0}$ *for constants* $C > 0$ *and* $0 < \tau_0 < \tau/4 - 1$, *then for any* $\theta > 0$, *there exist constants* $C_1 > 0$, $\eta_1 > 0$ *and* $\eta_2 > 0$ *such that, if* $\frac{\log p}{n} \leq \eta_2$ *and* λ *is chosen as in part (a), then*

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\max} \leq C_1 M_p \left(\sqrt{\frac{\log p}{n}} + M_p \frac{\sqrt{p}}{p_d} \right) \tag{3.2}$$

with probability at least

$$1 - O\left(p^{-\theta} + n^{-\tau/4+1+\tau_0} (\log p)^{\tau/4} \right).$$

Theorem 1 shows that the estimation error $\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0$ naturally decomposes into two components: the error from estimating $\boldsymbol{\Omega}_c$ based on compositional observations, and the error incurred when identifying $\boldsymbol{\Omega}_0$ from $\boldsymbol{\Omega}_c$. The quantity M_p affects both terms. If Condition 1 is further relaxed to require only $\max_{1 \leq j \leq p} E\{\exp(\eta Y_j^2)\} \leq K$ or $\max_{1 \leq j \leq p} E(|Y_j|^\tau) \leq K$, analogous results continue to hold, with the constants multiplying $\sqrt{(\log p)/n}$ in Inequalities (3.1)-(3.2) modified accordingly (Supplementary Text S3). To extend these guarantees to other matrix norms, we use the following inequality (proved in Supplementary Text S4).

Lemma 1. *If* $\|\boldsymbol{b}\|_1 \leq \|\boldsymbol{a}\|_1$ *for two vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$, *then for any* $C > 0$ *and* $0 \leq \alpha \leq 1$,

$$\|\boldsymbol{b} - \boldsymbol{a}\|_1 \leq 2 \left(C^{-\alpha} + C^{1-\alpha} \right) \|\boldsymbol{a}\|_\alpha^\alpha \|\boldsymbol{b} - \boldsymbol{a}\|_\infty^{1-\alpha}.$$

In particular, when $C = 1$,

$$\|\boldsymbol{b} - \boldsymbol{a}\|_1 \leq 4\|\boldsymbol{a}\|_\alpha^\alpha \|\boldsymbol{b} - \boldsymbol{a}\|_\infty^{1-\alpha}.$$

Combining Theorem 1 with Lemma 1 yields the following convergence rates for the Amcoda estimator (proved in Supplementary Text S5).

Theorem 2. *Suppose* $M_p = o\big(p_d\sqrt{\frac{\log p}{np}}\big)$. *Under the conditions of Theorem 1(a) (or 1(b)), the Amcoda estimator* $\widehat{\boldsymbol{\Omega}}$ *satisfies*

$$\begin{aligned}
\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\max} &= O\left(M_p\left(\frac{\log p}{n}\right)^{1/2}\right), \\
\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_1 &= O\left(s_0(p)M_p^{1-q}\left(\frac{\log p}{n}\right)^{(1-q)/2}\right), \\
\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_2 &= O\left(s_0(p)\sqrt{p}M_p^{1-q}\left(\frac{\log p}{n}\right)^{(1-q)/2}\right), \\
\frac{1}{p}\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\mathrm{F}}^2 &= O\left(s_0(p)M_p^{2-q}\left(\frac{\log p}{n}\right)^{(2-q)/2}\right)
\end{aligned}$$

with probability at least $1 - O\left(p^{-\theta}\right)$ *(or with the corresponding probability bound in Theorem 1(b)).*

To recover the support of $\boldsymbol{\Omega}_0$, let $\omega^* = \min_{\substack{1\leq k,l\leq p \\ \omega_{kl}^0 \neq 0}} |\omega_{kl}^0|$ denote the minimum signal strength. For a threshold $\tau_{np} > 0$, define the hard-thresholding estimator $\widehat{\boldsymbol{\Omega}}^\tau = (\widehat{\omega}_{kl}^\tau)$, where $\widehat{\omega}_{kl}^\tau = \widehat{\omega}_{kl} I(|\widehat{\omega}_{kl}| > \tau_{np})$. Successful recovery requires ω^* to exceed the estimation noise level. The following corollary shows that, under suitable conditions, the hard-thresholding estimator $\widehat{\boldsymbol{\Omega}}^\tau$

can recover the support set of $\mathbf{\Omega}_0$ with high probability (proved in Supplementary Text S6).

Corollary 1. *Under the assumptions of Theorem 1, if the threshold τ_{np} is chosen as the upper bound on the right-hand side of Inequality* (3.1) *(or* (3.2)*) and $\omega^* > 2\tau_{np}$, then the hard-thresholding estimator $\widehat{\mathbf{\Omega}}^{\tau}$ satisfies*

$$\operatorname{sgn}\left(\widehat{\mathbf{\Omega}}^{\tau}\right) = \operatorname{sgn}\left(\mathbf{\Omega}_0\right)$$

with probability at least $1 - O(p^{-\theta})$ (or with the corresponding probability bound in Theorem 1(b)).

4. Numerical Results

4.1 Simulation studies

We compare the numerical performance of Amcoda with three competing methods: gmcoda (Fang , 2023), glasso (Friedman, Hastie and Tibshirani , 2008) and SPIEC-EASI (Tipton et al. , 2018). We simulate n independent observations of $\boldsymbol{Y}$ from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Omega}_0^{-1}$. Throughout, we set $d = 2$ and generate two compositional vectors from $\boldsymbol{Y}$ using the relation in Equation (2.1). The mean vector $\boldsymbol{\mu}$ is drawn uniformly from $[-0.5, 0.5]^p$, and the precision matrix $\mathbf{\Omega}_0$ is constructed under four sparse network structures — band, hub, block and

random — following the procedure in Zhang, Wang and Lin (2025).

Our goal is to recover the conditional correlation network of the latent vector $\boldsymbol{Y}$ from the observed compositional data $\boldsymbol{X}$. Estimation accuracy is evaluated using the matrix ℓ_1, ℓ_2 and Frobenius norms of $\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0$, while edge recovery is assessed using the false positive rate (FPR), recall, precision and F1 score. Each experiment is repeated 20 times and results are summarized by averaging each metric over the replications. Computations are performed on a Windows Server system equipped with a Hygon C86 5380 16-core processor (2.50 GHz) and 128 GB of RAM.

The competing methods differ in both modeling assumptions and implementation details. Glasso is designed for absolute (unconstrained) data and therefore is applied after a log transformation of the compositional observations, whereas SPIEC-EASI relies on the centered log-ratio transformation. Amcoda selects tuning parameters via 5-fold cross-validation, while gmcoda uses the Bayesian information criterion. For edge recovery, glasso and SPIEC-EASI use stability selection (Liu, Roeder and Wasserman , 2010).

Table 1 reports results for the four network structures under both a large-sample regime ($n > p_1 + p_2$) and a high-dimensional regime ($n \leq p_1 + p_2$). Across all configurations, Amcoda achieves substantially smaller

estimation errors than the competing methods. Amcoda also attains higher recall, precision and F1 scores while maintaining a lower FPR in most settings. In contrast, gmcoda and SPIEC-EASI tend to exhibit higher FPR, resulting in excessive false edges, whereas glasso performs the worst overall — highlighting the importance of explicitly accounting for compositional constraints.

Amcoda is not the fastest method considered, but it provides the best overall performance. Its computational cost — driven primarily by cross-validation for the column-wise tuning parameters — can be substantially reduced via parallel computing. To illustrate this, we compare the running time of Amcoda and its competitors for the band network with $p = 100$, 200, 300, 400, 500 (Table S1). The parallel implementation of Amcoda using 20 threads is significantly faster than glasso and SPIEC-EASI, and remains computationally feasible in practice.

To assess performance with more than two compositional vectors, we conduct additional simulations with $p = 300$ and $d = 3,\ 4,\ 5$, where the total dimension is allocated approximately evenly across the d vectors. The results (Table S2) are consistent with those in Table 1, confirming that Amcoda performs robustly for multiple compositional vectors ($d \geq 2$). To further evaluate performance in a high-dimensional regime resembling real

microbiome studies (e.g., thousands of taxa), we also consider the simulation setting with $p = 1000$ and $d = 3$. Amcoda remains computationally feasible and maintains strong estimation accuracy in this setting (Table S3).

To study unbalanced cross-domain dimensions, simulations are conducted with ratios $p_1 : p_2 = 2, 4, 8, 10$ while keeping $p_1 + p_2 = 300$. The results (Table S4) show that Amcoda's performance is stable across these ratios, with only minor fluctuations across evaluation metrics.

To further demonstrate the advantage of Amcoda for multiple compositional vectors, we apply CARE — a method designed for a single compositional vector — directly to the multi-vector data. Across all scenarios, Amcoda substantially outperforms CARE in terms of precision, recall, and F1 score, highlighting the benefit of explicitly accounting for multiple simplex constraints (Table S5).

We also assess Amcoda's performance under varying levels of sparsity. Specifically, for the band network with $p = 100$ and $p = 300$, we generate count data from a multinomial distribution, Multinomial$(m_i, \boldsymbol{X}_i)$, where the total counts m_i are sampled uniformly from three ranges — $[20p, 20p + 5000]$, $[4p, 5p]$ and $[p, 2p]$ — yielding approximately 5%, 25% and 50% zeros, respectively. As expected, Amcoda's performance deteriorates as the proportion of zeros increasing (Table S6).

Finally, to assess robustness under more realistic abundance heterogeneity, we conduct simulations in which the mean vector $\boldsymbol{\mu}$ is drawn uniformly from the broader range $[0, 5]^p$, rather than the narrower interval $[-0.5, 0.5]^p$ used in Table 1. Amcoda continues to outperform competing methods in estimation error and F1 score, demonstrating robust performance under heterogeneous abundances (Table S7).

4.2 Real data analysis

Amcoda is applied to a gut microbiome dataset from Haak et al. (2021) to infer a microbial interaction network involving both bacteria and fungi. The dataset consists of fecal samples from 59 subjects, with bacterial abundances profiled by 16S rRNA sequencing and fungal abundances measured by ITS sequencing, yielding 180 bacterial and 18 fungal operational taxonomic units (OTUs). We remove OTUs present in fewer than 20% of samples and discard samples with more than 80% zero counts. After preprocessing, 47 samples remain, containing 33 bacterial and 15 fungal OTUs. For the remaining zeros, we add a pseudo-count of 0.5 to all counts and then convert the data to compositions.

Table 2 reports the numbers of inferred edges for Amcoda and the competing methods. Amcoda, gmcoda and SPIEC-EASI all detect cross-

Table 1: Performance comparisons of Amcoda and its competitors (gmcoda, glasso and SPIEC-EASI) for band, hub, block and random networks with $(n, p_1, p_2) = (200, 50, 50)$ and $(200, 150, 150)$. Results are averaged over 20 simulation replicates.

$p_1 = p_2$	Method	$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_1$	$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_2$	$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_{\mathrm{F}}$	FPR	Recall	Precision	F1	Time (s)
Band									
50	Amcoda	3.57	2.83	10.92	0.03	0.85	0.51	0.64	19
	gmcoda	3.79	3.39	14.74	0.07	0.80	0.31	0.45	1
	glasso	4.22	3.84	17.32	0.02	0.01	0.02	0.01	98
	SPIEC-EASI	3.84	3.35	14.40	0.03	0.76	0.54	0.63	35
150	Amcoda	3.75	3.23	23.19	0.01	0.68	0.63	0.65	316
	gmcoda	3.97	3.55	27.37	0.03	0.68	0.25	0.37	11
	glasso	4.45	3.89	29.75	0.02	0.01	0.01	0.01	4312
	SPIEC-EASI	3.99	3.32	24.05	0.01	0.81	0.49	0.61	641
Hub									
50	Amcoda	4.34	2.02	7.74	0.01	0.86	0.59	0.70	20
	gmcoda	6.51	3.09	13.78	0.02	0.48	0.38	0.43	1
	glasso	9.43	4.32	19.89	0.09	0.06	0.01	0.01	203
	SPIEC-EASI	7.18	3.56	17.81	0.05	0.64	0.21	0.32	59
150	Amcoda	4.92	2.24	13.48	0.01	0.86	0.50	0.63	318
	gmcoda	6.73	3.26	23.07	0.01	0.59	0.42	0.49	8
	glasso	8.01	4.11	35.79	0.00	0.00	0.04	0.01	4310
	SPIEC-EASI	6.28	3.44	28.73	0.01	0.91	0.35	0.51	768
Block									
50	Amcoda	5.10	3.00	10.08	0.02	0.76	0.58	0.66	18
	gmcoda	6.23	4.17	15.82	0.03	0.57	0.39	0.46	1
	glasso	7.58	4.86	20.52	0.17	0.21	0.04	0.07	120
	SPIEC-EASI	6.38	4.09	17.63	0.04	0.80	0.39	0.52	49
150	Amcoda	6.52	3.65	20.75	0.00	0.73	0.66	0.69	320
	gmcoda	7.36	4.38	29.09	0.02	0.66	0.35	0.45	14
	glasso	8.90	5.79	44.70	0.00	0.02	0.07	0.03	3654
	SPIEC-EASI	7.69	4.86	37.87	0.02	0.74	0.34	0.46	945
Random									
50	Amcoda	5.41	3.26	11.42	0.02	0.76	0.65	0.70	20
	gmcoda	8.01	4.42	17.30	0.04	0.59	0.36	0.44	1
	glasso	9.61	5.89	27.07	0.01	0.00	0.01	0.00	109
	SPIEC-EASI	9.39	5.63	25.82	0.09	0.39	0.15	0.21	85
150	Amcoda	6.33	3.41	21.73	0.01	0.73	0.64	0.68	339
	gmcoda	8.36	3.96	27.35	0.02	0.77	0.32	0.45	29
	glasso	9.18	6.02	48.03	0.02	0.04	0.04	0.04	3214
	SPIEC-EASI	8.53	5.37	43.82	0.03	0.63	0.25	0.35	1835

domain interactions — with 15, 81 and 1 edge(s), respectively — whereas glasso identifies none. Figure 1 shows the overlaps among inferred edges. Notably, all methods except glasso identify a negative association between *Agathobacter* (bacteria) and *Aspergillus* (fungi), consistent with findings from a COVID-19 microbiome study (Lv et al. , 2021). In addition, both Amcoda and gmcoda detect interactions between *Streptococcus* and *Candida*, in agreement with Sridhar et al. (2020).

To assess false discoveries for Amcoda, we permute the fungal samples to disrupt any genuine cross-domain associations. Across 20 permutations, the average numbers of false positives are 18 (Amcoda), 15 (gmcoda), 7 (glasso) and 6 (SPIEC-EASI). Although Amcoda yields most false positives, it also exhibits substantially higher power for edge detection in this dataset. To determine whether this tendency is specific to the gut microbiome dataset or reflects a more general pattern, we conduct an additional simulation under a null network (i.e., a true graph with no edges), using the same sample size and dimensions as in the real data ($n = 47$, $p_1 = 33$ and $p_2 = 15$). The results show that Amcoda tends to produce a higher false discovery count than the competing methods, consistent with the empirical findings (Figure S1). This suggests that there remains room for improvement in Amcoda's tuning-parameter selection procedure.

Table 2: Number of network edges inferred by Amcoda, gmcoda, glasso and SPIEC-EASI for the microbiome dataset from Haak et al. (2021).

Method	Bacteria-Bacteria	Fungi-Fungi	Bacteria-Fungi	Total
Amcoda	23	9	15	47
gmcoda	148	37	81	266
glasso	11	7	0	18
SPIEC-EASI	23	9	1	33

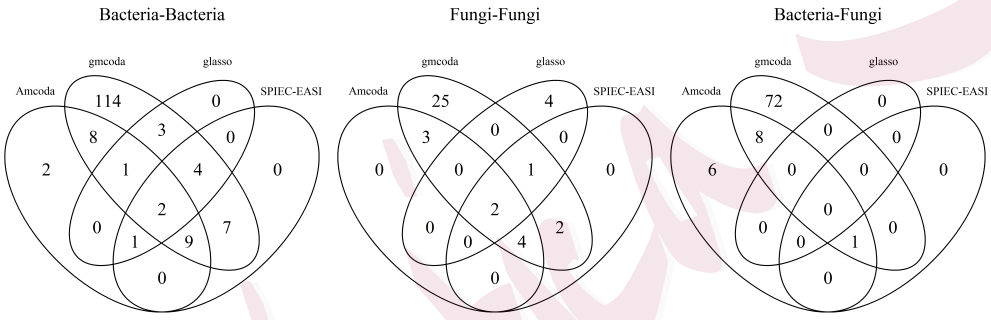

Figure 1: Overlap of network edges inferred by Amcoda, gmcoda, glasso and SPIEC-EASI on the microbiome dataset from Haak et al. (2021).

5. Conclusion

In this manuscript, we develop a new estimation framework for estimating the precision matrix of unobserved latent absolute abundances from observed multiple compositional vectors. We derive a closed-form expression for the Moore-Penrose inverse of the covariance matrix of the log-ratio-

transformed multiple compositional vectors, expressed directly in terms of the precision matrix of latent absolute abundances. Leveraging this relationship, we propose Amcoda, a constrained ℓ_1-minimization estimator, and establish convergence rates under various matrix norms as well as support recovery guarantees. Simulation studies show that Amcoda consistently outperforms existing methods in both estimation accuracy and edge recovery. An application to microbiome data further demonstrates its ability to identify biologically meaningful cross-domain interactions between bacteria and fungi.

Overall, Amcoda provides a principled extension of precision matrix estimation from a single compositional vector to the multi-vector setting, combining strong theoretical guarantees with robust empirical performance. Future work includes developing likelihood-based inference under multivariate normal models for latent absolute abundances and exploring likelihood-driven strategies for the tuning-parameter selection.

Supplementary Materials

Proofs of all theoretical results and additional simulation studies are provided in the Supplementary Material.

Acknowledgements

This work was supported by the Beijing Outstanding Young Scientist Program [JWZQ20240101027], the Beijing Natural Science Foundation [JR25003] and the National Natural Science Foundation of China [12101425, 12171328].

References

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

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

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

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Haak, B. W., Argelaguet, R., Kinsella, C. M., Kullberg, R. F. J., Lankelma, J. M., Deijs, M. and et al. (2021). Integrative transkingdom analysis of the gut microbiome in antibiotic perturbation and critical illness. *mSystems* **6**, e01148–20.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks.

PLoS Computational Biology **11**, e1004226.

Liang, W., Wu, Y. and Ma, X. (2022). Robust sparse precision matrix estimation for high-dimensional compositional data. *Statistics & Probability Letters* **184**, 109379.

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems* **24**, 1432–1440.

Liu, W. and Luo, X. (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis* **135**, 153–162.

Lv, L., Gu, S., Jiang, H., Yan, R. and Li, L. (2021). Gut mycobiota alterations in patients with COVID-19 and H1N1 infections and their associations with clinical features. *Communications Biology* **4**, 480.

Sridhar, S., Suprabha, B. S., Shenoy, R., Suman, E. and Rao, A. (2020). Association of Streptococcus mutans, Candida albicans and oral health practices with activity status of caries lesions among 5-year-old children with early childhood caries. *Oral Health and Preventive Dentistry* **18**, 911–919.

Tipton, L., Müller, C. L., Kurtz, Z. D. Huang, L., Kleerup, E., Morris, A. and et al. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* **6**, 1–14.

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

REFERENCES

Zhang, X. and He, F. (2019). Large precision matrix estimation for compositional data. *Statistics and Application* **8**, 777–783.

Shen Zhang

School of Statistics and Data Science, Qufu Normal University, China.

E-mail: zhangshen@qfnu.edu.cn

Huaying Fang

Beijing Advanced Innovation Center for Imaging Theory and Technology, Capital Normal University, China; Academy for Multidisciplinary Studies, Capital Normal University, China.

E-mail: hyfang@cnu.edu.cn

Tao Hu

School of Mathematical Sciences, Capital Normal University, China.

E-mail: hutao@cnu.edu.cn