EleMi: A robust method to infer soil ecological networks with better community structure

Nan Chen[0000−0003−4257−7265] and Doina Bucur[0000−0002−4830−7162]

University of Twente, 7522 NB Enschede, The Netherlands
n.chen@utwente.nl, d.bucur@utwente.nl

Abstract. Soil ecological networks enable us to better understand the complex interactions among a great number of organisms in soil. Soil communities are biotic groups with similar environmental and resource preferences. Community detection thus provides insights into the mechanisms of the soil ecosystem. Therefore, inferring ecological networks with clear community structure is essential for investigating the soil ecosystem. We propose *Elastic net regularized Multi-regression* (EleMi), a new method to infer soil ecological networks. To better find the community structure, EleMi does not infer pairwise interactions, but considers all organisms simultaneously. Specifically, it regresses the abundance of all other taxa to one taxon (with shared parameters across soil samples) and employs Elastic net to avoid over-sparsity and stochasticity. The results on both synthetic and real biotic data show that EleMi is more robust and can infer ecological networks with clearer community structure.

Keywords: Soil ecological networks · Community structure · Elastic Net

1 Introduction

The soil contains a tremendous number of organisms, including bacteria, fungi, nematodes, protists, etc. Even a handful of soil can contain millions to billions of organisms [17], most of which are microscopic. Furthermore, these organisms are not independent, but interact with each other in various ways [8]. It is challenging to model and study this intricate soil ecosystem.

Nowadays, sequencing technology for genetic information enables researchers to understand the *types* and *abundances* of organisms in the soil [12]. Network science can then investigate the complex interactions among them. *Ecological networks* (in our case, undirected co-occurrence networks) consist of nodes and edges, with nodes corresponding to organisms and edges to associations between them, estimated on the basis of their abundance. They have been an important tool for investigating soil ecosystems [5,20], and highly connected communities have been found in these networks [21,9]. The ideal method for the inference of ecological networks from data would (1) retain the *density* and *community structure* present in the real ecosystem, because this internal structure is crucial for studying the ecosystem in a modular way, and (2) be *robust* across soil types, datasets and experiments.

There already exist methods to construct ecological networks from organism abundance data, but these have limitations. Traditionally, Pearson and Spearman correlation methods [5,23] compute the pairwise correlation coefficient between organism abundance values across soil samples. These methods always result in *dense* (almost fully connected) networks with many spurious correlations, so in practice, the ecologists would employ extra thresholds to make the networks sparser. To overcome this, SparCC [7] estimates the correlations iteratively, to bolster the assumption of network *sparsity*, and to accommodate uncertainties arising from random sampling. This leads to high computational complexity. Unlike SparCC, CCLasso [6] accounts for sparsity with Lasso regularization, and estimates the correlation matrix more accurately. However, none of these methods account for the network community structure.

On the other hand, soil *communities* are biotic groups with similar environmental and resource preferences. The detection of community structure gives insights into interaction patterns, and thus into the mechanisms of the soil ecosystem—so modular ecological networks are desirable. However, Pearson, Spearman, SparCC, and CCLasso as methods of network inference all estimate interactions between organisms pairwise and thus overlook multi-organism interactions, leading to a limited and potentially biased understanding of the true complexity of the soil ecosystem. **SPIEC-EASI** [15] methods were then proposed, with two schemes (which we abbreviate here **SE_MB** and **SE_GL**) to infer sparse networks, both considering interactions between all organisms simultaneously. Both SE_MB and SE_GL methods perform well on accuracy and reproducibility, and provide a more nuanced understanding of the relationships between organisms. However, Lasso regression (an important part of these methods) works poorly under *multicollinearity*, which is always present in biotic data. Lasso has the propensity to select one at random from the multicollinear group [24], thus is more stochastic and also sparser. Ridge regression is a more robust option, but does not give a sparse solution. Instead, Elastic net [24] uses a linear combination of Lasso and Ridge, which allows us to combine their advantages [14].

In the present study, we develop a novel method called *Elastic net regularized Multi-regression* (**EleMi**) to infer ecological networks from abundance data. EleMi considers all organisms simultaneously by regressing the abundance data of all other organisms to one organism. To avoid excessive sparsity and stochasticity in the inferred network, EleMi employs Elastic net instead of Lasso. We compare our EleMi with all previous methods. We run this comparison with both *synthetic datasets* with different ground-truth community settings, and *real-world datasets* with different soil types and biological kingdoms. On synthetic data, when measuring the performance of organism-to-organism edge prediction, EleMi achieves an overall edge accuracy of 87.93 ± 0.75, and a precision of 31.14 ± 11.00 for the edges inside communities (out of a maximum of 100). In addition, it shows superior stability across different community structure settings: even when the community structure is weak, EleMi has the highest intra-community precision among all methods. On real-world datasets, EleMi obtains larger modularity (corrected Q) values on most network types. The result indicates that our proposed EleMi method is *more robust* and *more sensitive to community structure* than other methods. Besides, our method can infer networks with clearer community structure in different ecological scenarios.

2 Method

Notations and Problem Definition. In this paper, we use italics to indicate scalars, bold lowercase to represent vectors, bold uppercase to represent matrices. The l^1 norm of a matrix is denoted as $\|.\|_1$, the Frobenius norm as $\|.\|_F$, and the infinity norm as $\|.\|_\infty$. The transpose and inverse of matrices are $\mathbf{X}^T$ and $\mathbf{X}^{-1}$ respectively. $Diag(\mathbf{X})$ represents the diagonal matrix of $\mathbf{X}$. Besides, $\mathbf{x}_i$ and $\mathbf{x}^i$ denote the i_{th} row and column of $\mathbf{X}$.

Suppose that the abundance data of soil taxa is stored in matrix $\mathbf{D} \in \mathbb{R}^{n\times p}$, where n denotes the number of soil samples and p the number of taxa. We aim to infer the relationships between taxa to form an *undirected weighted network*. The network is represented as $\mathcal{G} = (V, E)$, where V is the vertex set of p taxa and E is the edge set. $\mathbf{A} \in \mathbb{R}^{p\times p}$ is the weighted adjacency matrix of network $\mathcal{G}$. We study the problem of inferring $\mathbf{A}$ from $\mathbf{D}$.

2.1 Data Normalization

The abundance data matrix $\mathbf{D}$ is obtained using sequencing technology, which has a series of limitations. First, the overall abundance counts, called *sequencing depths*, are artificially limited and differ among soil samples [19]. To address this, the sequencing depths of all soil samples are normalized to 1. Specifically, for the i_{th} row of $\mathbf{D}$, $\mathbf{d}_i^{'} = \mathbf{d}_i/\mathrm{sum}(\mathbf{d}_i)$. Second, this leads to a *compositional effect* since $\mathrm{sum}(\mathbf{d}_i^{'}) = 1$; an increase in abundance for one taxon in this sample would necessarily result in the abundance decrease of other taxa. This causes a negative bias, so prohibits statistical analyses among taxa. Thus, the *centered log-ratio* (clr) transformation [1] is typically used to remove the sum constraints of abundance data:

$$\mathbf{d}_i^{''} = \mathrm{clr}(\mathbf{d}_i^{'}) = \log(\mathbf{d}_i^{'}/\mathrm{geometric\_mean}(\mathbf{d}_i^{'}))\ . \tag{1}$$

Third, the abundance data has many zeroes (is sparse), arising from inefficient sequence sampling; this causes numerical problems for clr. A common practice is to replace zeros with small pseudo-counts [2,6,7,15]. For this, we use 1/10 of the minimum non-zero count in each soil sample. After all normalization steps, the abundance data matrix is transformed into the normalized abundance matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, which we use in the next regression step to infer the weighted adjacency matrix $\mathbf{A}$.

2.2 Elastic net regularized Multi-regression

Given the normalized abundance data matrix $\mathbf{X}$, we assume that if $\mathbf{x}^i$ can be regressed to $\mathbf{x}^j$, there is an association between taxa i and j. Instead of computing the associations pairwise, EleMi considers all taxa simultaneously. Specifically, it regresses the abundance of all other taxa to one taxon, with shared parameters across soil samples. The objective function is:

$$\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\|_F^2 + \mu_1\|\mathbf{A}\|_1 + \mu_2\|\mathbf{A}\|_F^2, \quad \text{s.t.}\, Diag(\mathbf{A}) = 0 \tag{2}$$

with μ_1 and μ_2 penalty parameters for the l^1 norm (Lasso) and F-norm (Ridge).

To solve (2) more easily and effectively, we employ the alternating direction multiplier method (ADMM) [3] to break down the original problem into two subproblems. Specifically, in our problem, by introducing $\mathbf{Z} \in \mathbb{R}^{p\times p}$ to replace $\mathbf{A}$ in the l^1 norm, the objective function becomes:

$$\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\|_F^2 + \mu_1\|\mathbf{Z}\|_1 + \mu_2\|\mathbf{A}\|_F^2, \quad \text{s.t.}\, \mathbf{Z} = \mathbf{A} \tag{3}$$

The augmented Lagrange function of (3) is defined as:

$$L(\mathbf{A},\mathbf{Z},\mathbf{Y},\rho) = \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\|_F^2 + \mu_1\|\mathbf{Z}\|_1 + \mu_2\|\mathbf{A}\|_F^2 + \frac{\rho}{2}\left\|\mathbf{Z}-\mathbf{A}+\frac{\mathbf{Y}}{\rho}\right\|_F^2 \tag{4}$$

where ρ is a positive penalty parameter, and $\mathbf{Y} \in \mathbb{R}^{p\times p}$ is the dual matrix (known as Lagrange multiplier). Then, the problem is broken down into two subproblems:

$$\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\|_F^2 + \mu_2\|\mathbf{A}\|_F^2 + \frac{\rho}{2}\left\|\mathbf{Z}-\mathbf{A}+\frac{\mathbf{Y}}{\rho}\right\|_F^2 \tag{5}$$

$$\min_{\mathbf{Z}} \mu_1\|\mathbf{Z}\|_1 + \frac{\rho}{2}\left\|\mathbf{Z}-\mathbf{A}+\frac{\mathbf{Y}}{\rho}\right\|_F^2 \tag{6}$$

These two subproblems can be optimized separately. Specifically, with $\mathbf{Z}$ fixed, the closed-form solution of (5) is:

$$\mathbf{A} = (2\mu_2\mathbf{I} + \mathbf{X}^T\mathbf{X} + \rho\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \rho\mathbf{Z} + \mathbf{Y}) \tag{7}$$

where $\mathbf{I}$ is the unit matrix. (6) can also be solved by soft thresholding [16]:

$$\mathbf{Z}_i^j = S_{\frac{2\mu_1}{\rho}}\left(\mathbf{A}_i^j - \frac{\mathbf{Y}_i^j}{\rho}\right) \tag{8}$$

where $S_\lambda(v)$ represents the shrinkage thresholding operator with the input value v and the threshold parameter λ:

$$S_\lambda(v) = \text{sign}(v)\cdot\max\left(0, |v| - \lambda\right) \tag{9}$$

where sign(v) is the sign function: $\text{sign}(v) = v/|v|$ if $v \neq 0$, otherwise $\text{sign}(v) = 0$. The dual matrix can be updated by:

$$\mathbf{Y} = \mathbf{Y} + \rho(\mathbf{Z} - \mathbf{A}) \tag{10}$$

The optimization steps are illustrated in Algorithm 1. The source code is publicly accessible on GitHub: https://github.com/nan-vince-chen/EleMi.

Algorithm 1 EleMi

```
Input: normalized abundance data matrix $\mathbf{X}$, $\mu_1$, $\mu_2$
Initialize: $\rho_0 = 0.1, \rho_{max} = 10^{10}, \beta_0 = 1.1, \mathbf{Y}_0 = \mathbf{Z}_0 = \mathbf{0}, t = 0, \text{threshold}_{\text{conv}} = 10^{-7}, \text{threshold}_\beta = 10^{-4}$
while $\|\mathbf{Z}_t - \mathbf{A}_t\| > \text{threshold}_{\text{conv}}$ do
    Fix $\mathbf{Z}_t$, compute $\mathbf{A}_{t+1}$ by (7)
    Fix $\mathbf{A}_{t+1}$, compute $\mathbf{Z}_{t+1}$ by (8)
    Compute $\mathbf{Y}_{t+1}$ by (10)
    Update $\rho$ by $\rho_{t+1} = \min(\rho_{max}, \beta\rho_t)$, where
        $$\beta = \begin{cases} \beta_0, & \text{if } \|\mathbf{Z}_{t+1} - \mathbf{Z}_t\|_\infty, \|\mathbf{A}_{t+1} - \mathbf{A}_t\|_\infty > \text{threshold}_\beta \\ 1, & \text{otherwise} \end{cases} \tag{11}$$
    update $t = t + 1$
end while
Output: $\mathbf{A} = (\mathbf{Z} + \mathbf{Z}^T)/2$
```

2.3 Datasets and Evaluation Metrics

The EleMi method is evaluated using both synthetic and real-world datasets. We describe these datasets below.

Synthetic Data. To measure performance on community structure and have the ground truth, we simulate the abundance data using a two-step pipeline:

1. First, the *adjacency matrix* is synthesized, with pre-defined communities, using the Gaussian probabilistic graph model [4]. In this model, nodes are partitioned into communities, and the connections between nodes are probabilistically determined using a Gaussian distribution based on their community assignment. The NetworkX implementation is used, with the following parameters: total number of nodes $p = 300$ (similar to the size of our real-world datasets described below), mean community size $s = 30$, and shape parameter $v = 5$ (which determines the variance of community size by s/v). We vary the probability of connections in- and outside communities ($p_{\text{in}} = \{0.35, 0.25, 0.15\}$, $p_{\text{out}} = 0.05$) to obtain communities with different density settings.

2. Second, the adjacency matrix is used to generate synthetic *abundance data* with the R package HARMONIES [13]. The process is repeated with different sample sizes $n = \{200, 1000\}$.

The network inference performance on synthetic data is evaluated with the accuracy (ACC), precision (PRE) and recall (REC) of the inferred edges. These performance metrics are also measured separately: *overall* (for all edges in the network), for *intra-community* edges, as well as for *inter-community* edges. We also repeat the community generation process randomly 10 times to evaluate stability.

The simulated adjacency matrices (ground truth) are actually unweighted. To make the comparison between our inferred weighted networks and this unweighted ground truth possible, we "binarize" our inferred weighted networks by applying a threshold on the weights. For this, we make the assumption that the weights of inferred edges (which can be positive or negative) represent, in their absolute value, the strength of connections (0 means no connection). Notably, most of the competing methods (except Pearson and Spearman) have already incorporated sparsity settings in their theoretical frameworks. In light of this, we establish a threshold of 0.8 for Pearson and Spearman networks additionally for better comparison.

Real-world Data. We also evaluate the performance on real-world datasets (collected in 2021 from Dutch soils): taxa from 4 different kingdoms (bacteria, fungi, nematodes, and protists) on 2 different soil types (sand and clay). Bacterial species in the soil are numerous (many thousands), so here they are aggregated to the genus level. For all other organisms, a network node represents a species. The abundance data per kingdom is obtained by different measurement processes, so we treat the kingdoms as separate ecological networks. Their sizes (number of taxa p per kingdom) ranges between 83 and 717.

Since there is no ground truth for the real-world ecological networks, we evaluate the network inference by using also the weighted modularity Q [18], defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left(\mathbf{A}_i^j - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \qquad (12)$$

where $\mathbf{A}_i^j$ represents the weighted adjacency matrix element, k_i and k_j are the weighted degrees of nodes i and j respectively, m is the total sum of edge weights, $\delta(c_i, c_j)$ is 1 if $c_i = c_j$ (nodes i and j are in the same community) and 0 otherwise. The community detection is implemented using *greedy_ modularity_ communities* function in NetworkX. However, this classical modularity Q has biases. One bias is towards the number of communities. To mitigate this, Yin et. al introduced eQ [22]. But, to compare Q values between different networks, bias towards network density also needs to be corrected, since a lower density would intrinsically lead to a bigger Q regardless of the community structure. To achieve this, we present the corrected Q (cQ) based on eQ, defined as:

$$cQ = Q \cdot \frac{|C|+1}{|C|} \cdot \frac{|E|/\epsilon - 1}{|E|/\epsilon} \tag{13}$$

where $|C|$ is the number of communities, $|E|$ is the number of edges, $\epsilon = 0.01 \times |V|^2$ is a scale parameter, where $|V|$ is the number of nodes.

3 Results and Discussion

Tuning parameters. The penalty parameters in EleMi, μ_1 and μ_2, were first tuned (to obtain the largest cQ) in the range $\{10^{-4}, 10^{-3}, ..., 10^3, 10^4\}$. Fig. 1 presents these results on two different real soil ecological networks; the results are consistent on all other network types. The figure shows how cQ varies with different combinations of values for μ_1 and μ_2. We observe that EleMi performs best in modularity when $\mu_1 = 10^{-1}$ and over a wide range of μ_2 (from 10^{-4} to 10^{-1}). We then use these tuned values for the evaluation.

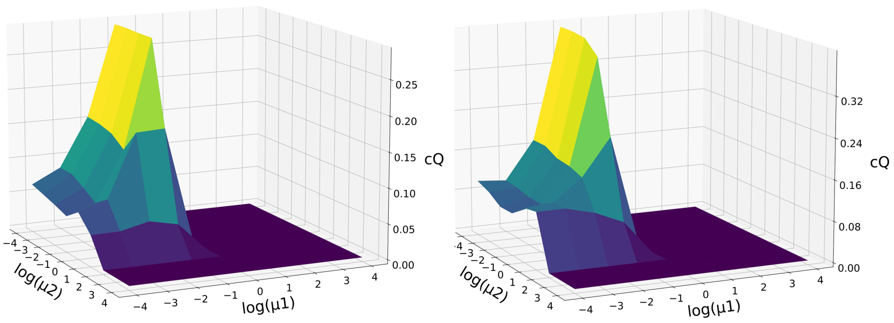

Fig. 1. cQ values vs μ_1 and μ_2 on two real soil ecological networks.

3.1 Results on synthetic data

Consistent with the settings, 10.40 ± 0.66 communities are generated, each containing 28.85 ± 7.01 nodes. To curtail redundant calculations, we fixed μ_1 and μ_2 as 10^{-1} and 10^{-2} in comparisons on synthetic data.

Fig. 2 shows ACC and PRE achieved by all 9 methods of network inference on synthetic datasets. Pearson, Spearman, and SparCC obtained poor overall ACC (considering all different settings in Fig. 2, Pearson: 23.43 ± 5.99; Spearman: 18.18 ± 3.32; SparCC: 14.55 ± 2.93). This is unacceptable. These methods yield dense (almost fully connected) networks with many spurious correlations [11], far from reality. The thresholded Pearson and Spearman obtain much higher

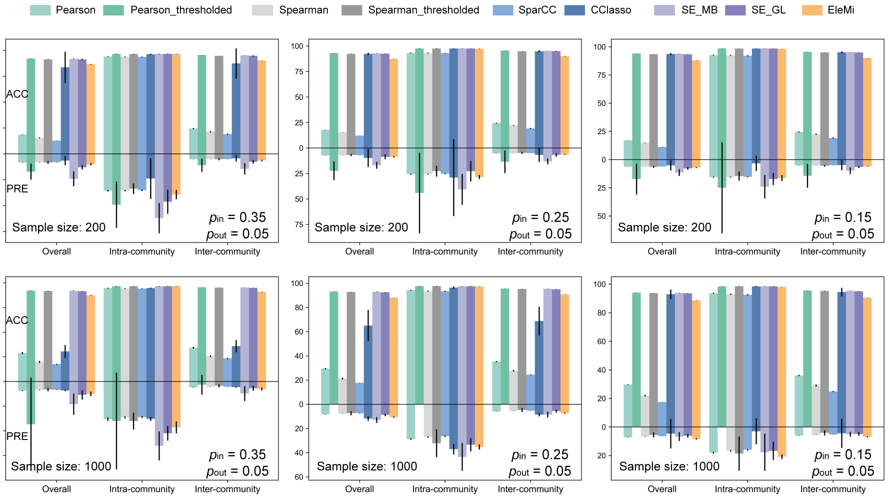

Fig. 2. Accuracy (ACC) and precision (PRE) on synthetic data with different settings. p_{in} and p_{out} represent connection probabilities in- and outside communities.

ACC by excluding part of those spurious edges. However, the effect of thresholding on PRE varies among Pearson and Spearman in different settings. When the sample size is limited (on the first row of Fig. 2, the sample size of 200 is smaller than the number of nodes), the thresholded Pearson gets higher PRE but is more unstable, while no significant difference is found between Spearman and thresholded Spearman. The standard deviations of thresholded methods keep increasing when the clarity of community structure decreases, which indicates that it is harder for thresholding to exclude spurious edges when the community structure is unclear. CClasso has better overall ACC (80.85 ± 22.50), but it is unstable: the complexity of recovering inter-community relationships hampers the accuracy of the inference process, leading to fluctuations in the overall network inference accuracy.

SE_MB, SE_GL, and our proposed EleMi method all have good overall ACC (considering all different settings in Fig. 2, SE_MB: 93.20 ± 0.96; SE_GL: 92.79 ± 1.05; EleMi: 87.93 ± 0.75). However, overall REC for SE_MB (0.74 ± 0.59) and SE_GL (1.07 ± 0.71) are much lower than REC for our method (8.98 ± 1.04), which shows that SPIEC-EASI methods might be over-sparse and wrongly exclude edges. Also, the intra-community overall PRE of SPIEC-EASI methods (considering all different settings in Fig. 2, SE_MB: 42.21 ± 22.11; SE_GL: 31.51 ± 16.07) and EleMi (31.14 ± 11.00) are higher than for the other methods (from Pearson: 26.97 ± 8.36 to CClasso: 22.38 ± 21.14).

It is worth noting that when p_{in} is close to p_{out} (see the third column of Fig. 2), which means the community structure is not discernible, our method achieves a better intra-community PRE (bars in the middle, 20.84 ± 1.57) than

SPIEC-EASI methods (SE_MB: 17.70 ± 13.02; SE_GL: 16.73 ± 6.42), which indicates that EleMi is more sensitive to the community structure.

Robustness. SE_MB and SE_GL have significantly larger standard deviations of performance compared to EleMi, possibly due to their Lasso terms. Lasso terms are included in the competing methods to increase network sparsity, aiming to reduce spurious connections caused by compositional effects. However, multicollinearity is common and implies connections between nodes. When multicollinearity exists, Lasso tends to randomly select one variable rather than consider them all. This random selection process results in larger standard deviations and lower REC, especially for SPIEC-EASI. In summary, while Lasso mitigates compositional effects, it also weakens the inherent multicollinearity among nodes, which is undesirable for network inference tasks. Our method introduces an additional Ridge term to balance these two effects with the parameters μ_1 and μ_2. Unlike Lasso, Ridge considers multicollinear variables simultaneously, but assigns them smaller weights. This makes EleMi more stable and with a higher recall (REC) than competing methods. Also, as shown in Fig. 1, μ_1 has a greater influence on results compared with μ_2, which may indicate that mitigating the compositional effect is much more important than preserving the multicollinearity between nodes. This is also the reason why methods without Ridge terms also work well.

Community structure. We further investigate the modularity values Q and cQ on synthetic data. Table 1 shows the result with $p_{in} = 0.25, p_{out} = 0.05$, for sample size 200. EleMi obtains the closest density and number of communities to the ground truth. Although SE_MB and CClasso achieved higher Q and cQ values than EleMi (and also much higher than the ground truth), their communities are (a) unreasonably many and (b) extremely fragmented. Similar phenomena are also observed in thresholded Pearson and Spearman. In line with the previous findings, thresholding has a limited effect on the performance of Spearman, only making its communities more fragmented. Overall, thresholding does exhibit a potential benefit in excluding some spurious connections of Pearson (although it is not consistently stable). This may be useful in differential analyses, but may not be suitable for community detection tasks. Besides, also consistent with the previous results, the standard deviation of EleMi is smaller than that of CClasso and SPIEC-EASI methods.

3.2 Results on real data

Table 2 shows the performance of the different methods on real-world soil data. Our method exhibits stable and good performance on both Q and cQ, which indicates that it can infer networks with clear community structure and is more robust across different network types. It has been reported that Q values are normally between 0.3 to 0.7 and values above 0.7 are rare in practice [10]. Larger Q values do not always indicate better performance, because Q values have biases towards both the number of communities and network density. SE_MB and

Table 1. Comparison of Q and corrected Q (cQ) values on synthetic data with $p_{in} = 0.25, p_{out} = 0.05$, for sample size 200. $|C|$ is the number of communities and std is the standard deviation.

Methods	Q		cQ		density		$\|C\|$	
	mean	std	mean	std	mean	std	mean	std
Pearson	0.061	0.002	0.069	0.003	0.440	0.001	6.300	1.187
Pearson_thresholded	0.898	0.044	0.270	0.013	0.001	0.000	273.900	7.034
Spearman	0.168	0.005	0.244	0.010	0.453	0.001	2.100	0.300
Spearman_thresholded	0.076	0.006	0.023	0.002	0.005	0.000	252.600	2.538
SparCC	0.035	0.001	0.042	0.002	0.417	0.001	4.100	0.300
CClasso	0.454	0.282	0.139	0.084	0.002	0.002	149.400	138.958
SE_MB	0.537	0.075	0.162	0.023	0.002	0.001	135.800	46.927
SE_GL	0.158	0.047	0.048	0.014	0.004	0.002	178.300	48.019
EleMi	0.173	0.007	0.131	0.005	0.034	0.000	15.500	1.746
Ground truth	0.251	0.014	0.196	0.013	0.035	0.001	10.400	0.663

SE_GL achieved Q values around 0.8: this is attributed to their over-division into an unrealistically large number of communities given the number of taxa, as well as to the inherent sparsity of the network. Unlike SPIEC-EASI methods, our method infers networks with a more reasonable number of communities.

We also visualize the detected communities in ecological networks inferred for nematodes on clay soils (the network denoted N_C in Table 2). As shown in Fig. 3, SPIEC-EASI methods have many isolated nodes, and thus unreasonable numbers of communities detected by SE_MB (37 communities) and SE_GL (63 communities), despite larger uncorrected Q values. In contrast, EleMi has a clearer community structure for almost all nodes.

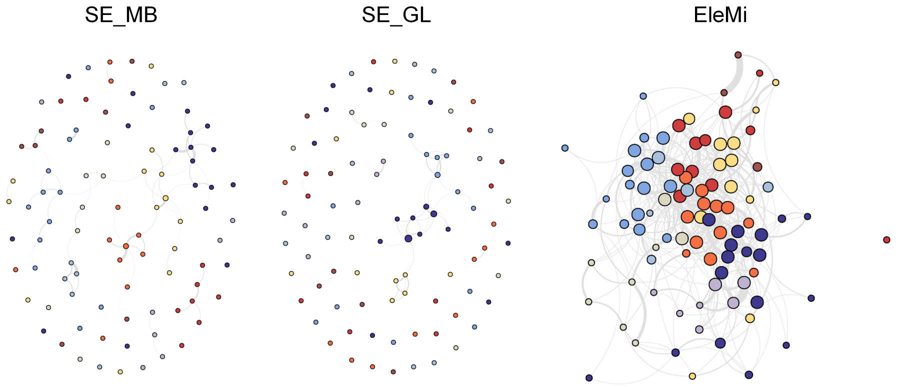

Fig. 3. Force-directed visualization of communities in ecological networks of nematodes on clay (N_C). Different colors represent different communities. The size of the nodes is proportional to their strength.

Table 2. Comparison of Q and corrected Q (cQ) on real-world soil ecological networks. The network names in the first column denote different organism kingdoms (B: Bacteria, F: Fungi, N: Nematodes, P: Protists) on different soil types (C: Clay, S: Sand). E.g., B_C is Bacteria on Clay. p is the number of nodes. $\lvert C\rvert$ is the number of communities.

Network (p)	Pearson				Pearson_thresholded				Spearman			
	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$
B_C (317)	0.081	0.095	0.461	5	0.833	0.251	0.002	208	0.093	0.137	0.482	2
B_S (298)	0.088	0.107	0.478	4	0.581	0.177	0.010	72	0.144	0.212	0.484	2
F_C (717)	0.111	0.130	0.439	5	0.875	0.265	0.004	111	0.020	0.029	0.491	2
F_S (660)	0.116	0.136	0.470	5	0.758	0.237	0.019	23	0.040	0.052	0.490	3
N_C (92)	0.151	0.169	0.458	7	0.864	0.263	0.004	65	0.063	0.093	0.481	2
N_S (83)	0.221	0.252	0.471	6	0.719	0.372	0.020	25	0.049	0.073	0.487	2
P_C (395)	0.083	0.094	0.453	6	0.901	0.271	0.001	254	0.035	0.052	0.485	2
P_S (348)	0.079	0.096	0.471	4	0.836	0.253	0.004	107	0.073	0.095	0.485	3
Network (p)	Spearman_thresholded				SparCC				CClasso			
	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$
B_C (317)	0.040	0.036	0.086	114	0.094	0.138	0.457	2	0.096	0.117	0.392	4
B_S (298)	0.123	0.110	0.083	45	0.089	0.116	0.465	3	0.125	0.162	0.364	3
F_C (717)	0.013	0.013	0.235	134	0.062	0.081	0.442	3	0.135	0.160	0.196	4
F_S (660)	0.047	0.044	0.165	68	0.068	0.089	0.462	3	0.484	0.359	0.029	8
N_C (92)	0.031	0.029	0.129	32	0.106	0.139	0.448	3	0.157	0.182	0.297	5
N_S (83)	0.059	0.059	0.165	19	0.130	0.169	0.452	3	0.618	0.504	0.040	11
P_C (395)	0.021	0.019	0.132	136	0.089	0.116	0.457	3	0.105	0.136	0.358	3
P_S (348)	0.044	0.041	0.105	103	0.067	0.082	0.457	4	0.674	0.204	0.002	101
Network (p)	SE_MB				SE_GL				**EleMi**			
	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$	Q	cQ	density	$\lvert C\rvert$
B_C (317)	0.588	0.186	0.009	18	0.207	0.063	0.014	128	0.436	0.288	0.025	9
B_S (298)	0.733	0.228	0.006	28	0.510	0.155	0.009	93	0.494	0.299	0.022	10
F_C (717)	0.520	0.165	0.007	18	0.092	0.109	0.166	4	0.602	0.193	0.011	15
F_S (660)	0.497	0.160	0.008	14	0.662	0.203	0.009	46	0.486	0.155	0.012	16
N_C (92)	0.809	0.249	0.007	37	0.622	0.190	0.005	63	0.439	0.398	0.054	9
N_S (83)	0.814	0.259	0.011	17	0.842	0.258	0.004	47	0.477	0.429	0.052	9
P_C (395)	0.536	0.170	0.008	18	0.075	0.052	0.031	159	0.424	0.180	0.017	13
P_S (348)	0.722	0.222	0.004	37	0.699	0.211	0.003	140	0.443	0.227	0.019	12

cQ further corrects Q. As shown in Fig. 4 in the trend lines, cQ mitigates the bias of Q values towards network density. Unlike eQ, cQ cannot serve as an objective metric for community detection, since the density of networks is already determined when detecting the communities. Instead, cQ allows for fairer comparisons of community structure between different networks.

Limitations. Firstly, the symmetry of adjacency matrix $\mathbf{A}$ in our method is guaranteed by computing the average of its transposition and itself, which is a sub-optimal solution. In the future, we would add symmetry as a condition in the theory without greatly increasing the computation complexity. Secondly, the synthetic adjacency matrix generated with predefined community structure should ideally be weighted, and the simulation of abundance data from a weighted adjacency matrix should also be better explored. We expect to implement these two steps into a future improved method. Thirdly, we only choose a traditional community detection method in order to show that EleMi-inferred networks can obtain good community structure even with the simplest community detection algorithm. Performance with other community detection methods or other quality metrics can be further investigated in the future.

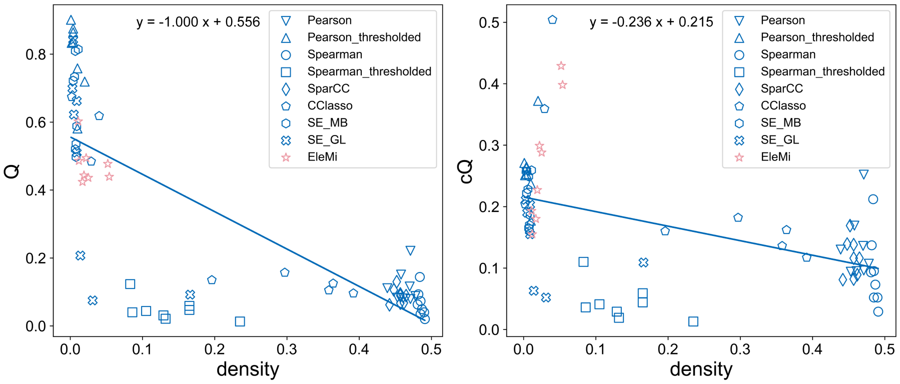

Fig. 4. Q and corrected Q (cQ) vs. network density. Data points represent different network-inference methods on real data. The trend lines show linear regression.

4 Conclusion

This paper proposes a robust method to infer soil ecological networks. In order to evaluate the proposed new method, we conduct comparison experiments on both synthetic and real-world datasets. On real-world datasets, rather than using the traditional Q value, we propose a new corrected Q (cQ) value to compare the quality of detected communities from networks of different sizes.

The results show that our method can infer ecological networks with stable performance under different community structure settings and is more sensitive to community structure. It is more robust when inferring networks with clearer community structure for different real-world ecological networks. Moreover, our proposed cQ can mitigate the bias of Q values towards network density thus allowing for fairer comparisons of community structure between different networks.

Acknowledgments. This study was funded by Dutch Research Council (NWO) (grant number P20-45) under the project SoilProS (Soil biodiversity analysis for sustainable production systems).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aitchison, J.: The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Methodol. **44**(2), 139–160 (1982). https://doi.org/10.1111/j.2517-6161.1982.tb01195.x
2. Ban, Y., An, L., Jiang, H.: Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinform. **31**(20), 3322–3329 (2015). https://doi.org/10.1093/bioinformatics/btv364

3. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011). https://doi.org/10.1561/2200000016
4. Brandes, U., Gaertler, M., Wagner, D.: Experiments on graph clustering algorithms. In: European Symposium on Algorithms. pp. 568–579. Springer (2003)
5. Delgado-Baquerizo, M., Reith, F., Dennis, P.G., Hamonts, K., Powell, J.R., Young, A., Singh, B.K., Bissett, A.: Ecological drivers of soil microbial diversity and soil biological networks in the southern hemisphere. Ecology **99**(3), 583–596 (2018). https://doi.org/10.1002/ecy.2137
6. Fang, H., Huang, C., Zhao, H., Deng, M.: CCLasso: correlation inference for compositional data through Lasso. Bioinform. **31**(19), 3172–3180 (2015). https://doi.org/10.1093/bioinformatics/btv349
7. Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. PLoS Comput. Biol. **8**(9), e1002687 (2012). https://doi.org/10.1371/journal.pcbi.1002687
8. Geisen, S., Briones, M.J., Gan, H., et al.: A methodological framework to embrace soil biodiversity. Soil Biol. Biochem. **136**, 107536 (2019). https://doi.org/10.1016/j.soilbio.2019.107536
9. Guo, Y., Wu, J., Yu, Y.: Differential response of soil microbial community structure in coal mining areas during different ecological restoration processes. Processes **10**(10), 2013 (2022). https://doi.org/10.3390/pr10102013
10. Gustafsson, M., Hörnquist, M., Lombardi, A.: Comparison and validation of community structures in complex networks. Phys. A Stat. Mech. Appl. **367**, 559–576 (2006). https://doi.org/10.1016/j.physa.2005.12.017
11. Hirano, H., Takemoto, K.: Difficulty in inferring microbial community structure based on co-occurrence network approaches. BMC Bioinform. **20**(1), 1–14 (2019). https://doi.org/10.1186/s12859-019-2915-1
12. Hirsch, P.R., Mauchline, T.H., Clark, I.M.: Culture-independent molecular techniques for soil microbial ecology. Soil Biol. Biochem. **42**(6), 878–887 (2010). https://doi.org/10.1016/j.soilbio.2010.02.019
13. Jiang, S., Xiao, G., Koh, A.Y., Chen, Y., Yao, B., Li, Q., Zhan, X.: HARMONIES: a hybrid approach for microbiome networks inference via exploiting sparsity. Front. Genet. **11**, 445 (2020). https://doi.org/10.3389/fgene.2020.00445
14. Katrutsa, A., Strijov, V.: Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst. Appl. **76**, 1–11 (2017). https://doi.org/10.1016/j.eswa.2017.01.048
15. Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput. Biol. **11**(5), e1004226 (2015). https://doi.org/10.1371/journal.pcbi.1004226
16. Loris, I., Verhoeven, C.: On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Problems **27**(12), 125007 (2011). https://doi.org/10.1088/0266-5611/27/12/125007
17. Luan, L., Jiang, Y., Cheng, M., Dini-Andreote, F., Sui, Y., Xu, Q., Geisen, S., Sun, B.: Organism body size structures the soil microbial and nematode community assembly at a continental and global scale. Nat. Commun. **11**(1), 6406 (2020). https://doi.org/10.1038/s41467-020-20271-4
18. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A. **103**(23), 8577–8582 (2006). https://doi.org/10.1073/pnas.060160210

19. Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., Wei, X.: A review of normalization and differential abundance methods for microbiome counts data. WIREs Comput. Stat. **15**(1), e1586 (2023). https://doi.org/10.1002/wics.1586
20. Wagg, C., Schlaeppi, K., Banerjee, S., Kuramae, E.E., van der Heijden, M.G.: Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. Nat. Commun. **10**(1), 4841 (2019). https://doi.org/10.1038/s41467-019-12798-y
21. Xue, L., Ren, H., Brodribb, T.J., Wang, J., Yao, X., Li, S.: Long term effects of management practice intensification on soil microbial community structure and co-occurrence network in a non-timber plantation. For. Ecol. Manage. **459**, 117805 (2020). https://doi.org/10.1016/j.foreco.2019.117805
22. Yin, C., Zhu, S., Chen, H., Zhang, B., David, B.: A method for community detection of complex networks based on hierarchical clustering. Int. J. Distrib. Sens. Netw. **11**(6), 849140 (2015). https://doi.org/10.1155/2015/849140
23. Zhang, W.: Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biol. **1**(2), 81 (2011). https://doi.org/10.0000/issn-2220-8879-networkbiology-2011-v1-0008
24. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**(2), 301–320 (2005). https://doi.org/10.1111/j.1467-9868.2005.00503.x