BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–4

Supplementary Material for the Bioinformatics submission "CCLasso: Correlation Inference for Compositional Data through Lasso"

Huaying Fang [1,2,3], Chengcheng Huang [4], Hongyu Zhao [5], and Minghua Deng [1,3,6,*]

[1]LMAN, School of Mathematical Sciences, Peking University, Beijing 100871, China.
[2]Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China.
[3]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China.
[4]College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China.
[5]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA.
[6]Center for Statistical Science, Peking University, Beijing 100871, China.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

This supplementary material includes the details about the identification condition for $\Sigma_{\ln y}$ and more analysis for both simulation studies and real data. In the simulation studies, we show that SparCC is robust for the tuning parameter choice, the consistent accuracy and reproducibility are not good measurements to judge the effect of estimated methods, and CCREPE's performance is similar to SparCC. In the real data analysis, we first add more detailed analysis including time comparison, shape explore through degree distribution, influence of reproducibility for the choice of top edges, the correlation of the estimated result between CCLasso and SparCC for HMP datasets. Then we use CCLasso and SparCC to estimate the interaction network of 15 microbes from an acid mine drainage environment. We find that SparCC gets too many nonsense edges and CCLasso can get more informational connections than SparCC.

1 IDENTIFICATION CONDITION

We'll proof that there is at most one sparse network $\Sigma_{\ln y}$ whose edge density is no greater than $\frac{1}{2} - \frac{1}{p-1}$ corresponding to the same $\Sigma_{\ln x}$, and this sparse density condition can not be relaxed.

Suppose there are 2 sparse matrices Σ_1 and Σ_2 for $\Sigma_{\ln y}$ corresponding to the same $\Sigma_{\ln x}$ and assume the number of nonzero entries in the lower triangle part of both these two sparse matrces is less than s. Then we have

$$\Sigma_{\ln x} = \Sigma_1 - a_1\mathbf{1}^T - \mathbf{1}a_1^T = \Sigma_2 - a_2\mathbf{1}^T - \mathbf{1}a_2^T.$$

So

$$\Sigma_1 - \Sigma_2 = (a_1 - a_2)\mathbf{1}^T + \mathbf{1}(a_1 - a_2)^T.$$

We can find that if $a_1 = a_2$ then $\Sigma_1 = \Sigma_2$. Since there is at most $2s$ nonzero in the lower triangle part, then $a_1 - a_2 = 0$ is true if and only if

$$\frac{p(p-1)}{2} - 2s \geq p.$$

So the sparse degree s should be satisfied

$$s \leq \frac{p(p-3)}{4},$$

and combined with the total possible edge number $\frac{p(p-1)}{2}$, we can get that if the edge density is no greater than $\frac{1}{2} - \frac{1}{p-1}$, then there is only one possible $\Sigma_{\ln y}$ corresponding to $\Sigma_{\ln x}$ and this sparse density condition can not be relaxed.

2 SIMULATION STUDIES

2.1 Robust of tuning parameter's choice for SparCC

It's not feasible to learn the tuning parameters of SparCC from cross validation of the data since there is no loss function in this method. But SparCC's result is not sensitive for the tuning parameter's choice since only one strongest correlated pair larger than given threshold α is removed in each iteration. Both the simulation studies and HMP datasets use the default threshold parameter $\alpha = 0.1$. Fig. S1 is the ROC curves for SparCC with different tuning parameters $\alpha = 0.05,\ 0.1,\ 0.5$ for different correlation structures. SparCC is robust for the tuning parameter α from the simulation result.

2.2 Consistent accuracy and reproducibility for simulation data

Since there is no true answer for the real data, there is not a good way to compare the performance between CCLasso and SparCC as the consistent accuracy and reproducibility may not be a nice criteria when the results from a subset of samples

*to whom correspondence should be addressed

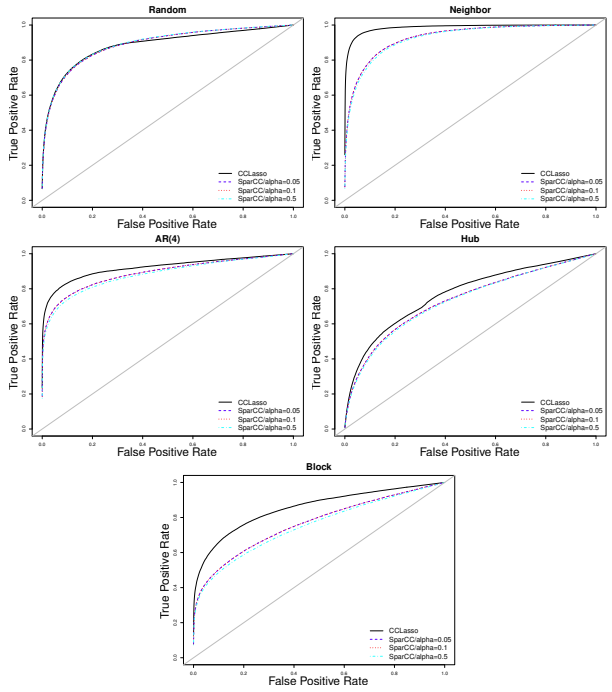

Fig. S1. ROC curves of SparCC with different tuning parameters (sample size is fixed as 300). The true positive rate is averaged over 100 replications after fixing the false positive rate and the gray line is the baseline reference.

Table S1. Consistent accuracy and reproducibility for CCLasso and SparCC from simulation data of the block model. (The results are averages over 20 replication runs with standard deviations in brackets.)

Sample Size	Accuracy		Reproducibility	
	CCLasso	SparCC	CCLasso	SparCC
100	2.92(0.00)	4.96(0.17)	0.80(0.31)	0.60(0.02)
200	2.19(0.06)	3.43(0.13)	0.65(0.02)	0.65(0.02)
300	2.01(0.05)	2.80(0.10)	0.69(0.02)	0.68(0.02)
500	1.59(0.07)	2.17(0.06)	0.75(0.02)	0.73(0.02)

is compared to all samples. We compare CCLasso and SparCC through the consistent accuracy and reproducibility for simulation data. The block model is used to explain the performance of CCLasso and SparCC (Table S1). We can find that the difference of the reproducibility of CCLasso and SparCC is negligible and the consistent accuracy is almost the same for these two methods. But from the main text we know CCLasso works better than SparCC in general from the ROC and the distance between estimated correlation matrix and the true one.

2.3 CCREPE is similar to SparCC from simulation studies

CCREPE is based on the distribution comparison between permutation and bootstrap to infer the significance of association for compositional data. The permutation and bootstrap are very common to infer the significance in statistics. We cannot get an

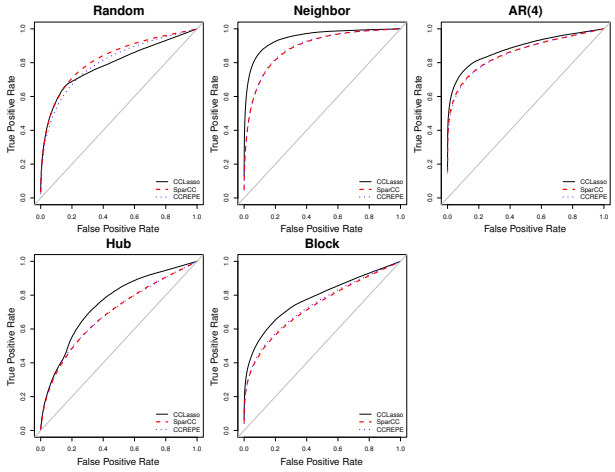

Fig. S2. ROC curves of CCREPE and CCLasso/SParCC with sample size equals 200.

available correlation strength measurement among the components from CCREPE but only a significant p-value. So we only compare the ROC curves between CCREPE and CCLasso/SparCC in Fig. S2 for simulations studies. We find that the ROC curve of CCREPE is similar to SparCC in the simulation studies. (Since the permutation and bootstrap of CCREPE are memory consumption and our PC is out of memory for large sample size, only sample size 200 is used to compare the performance among these three algorithms.)

3 REAL DATA

3.1 HMP datasets

3.1.1 Time comparison between CCLasso and SparCC for HMP datasets SparCC is much faster than CCLasso in general since there are several optimization procedures from the cross validation of CCLasso. We think the run time for CCLasso is acceptable for real problems such as HMP datasets. Table S2 show the run time comparison between CCLasso and SparCC for HMP datasets (PC: Intel(R) Core(TM) i5-2400 CPU, 4 GB MEM). SparCC is faster of the magnitude order of 2 than CCLasso in general. But the run time of CCLasso is in the acceptable range.

3.1.2 Shape explore of microbial correlation network for HMP datasets It is difficult to determine how natural networks are shaped since there is no clear boundary for network shapes. We explore the shapes of microbial correlation networks for HMP through the degree distribution. Fig. S3 shows the degree distributions of inferred correlation networks through CCLasso and SparCC with a common threshold 0.2. We can find there are more isolated nodes for CCLasso than SparCC in general.

3.1.3 Robust of reproducibility for top edges' choice There are negligible changes of the HMP datasets evaluation for using other top edges such as 10% and 40% while 25% is used in the main text. Table S3 shows the reproducibility for CCLasso and SparCC with different top edges 10%, 40%. We can find the reproducibility is robust for the top edges' choice.

Table S2. Run time (seconds) of CCLasso and SparCC for HMP datasets. (The results are averages over 10 replication runs with standard deviations in brackets.)

Sample Size	No. OTU	CCLasso	SparCC
152	34	18.73(0.09)	0.26(0.04)
193	25	20.45(0.25)	0.24(0.01)
196	36	17.29(0.35)	0.31(0.02)
197	36	28.37(0.27)	0.29(0.01)
51	53	22.09(0.10)	0.27(0.01)
123	31	9.29(0.07)	0.21(0.02)
45	21	19.84(0.15)	0.11(0.01)
203	38	63.29(0.28)	0.31(0.00)
22	15	36.14(0.16)	0.07(0.01)
54	54	14.93(0.08)	0.27(0.01)
85	38	10.22(0.22)	0.21(0.01)
184	41	29.58(0.49)	0.32(0.02)
190	32	12.95(0.04)	0.27(0.01)
205	41	118.60(1.28)	0.33(0.01)
207	39	65.51(0.38)	0.32(0.01)
197	39	46.45(0.13)	0.32(0.02)
207	35	70.88(0.45)	0.30(0.01)
52	26	14.99(0.05)	0.13(0.01)

Table S3. Reproducibility for CCLasso and SparCC of different body sites from HMP data with different top edges. (The results are the averages over 20 replication runs with standard deviations in brackets.)

Body Site	Sample Size	Reproducibility (top 10%)		Reproducibility (top 40%)	
		CCLasso	SparCC	CCLasso	SparCC
AntNar	152	0.68(0.07)	0.66(0.05)	0.70(0.03)	0.72(0.04)
AKerGin	193	0.70(0.06)	0.60 (0.06)	0.81 (0.03)	0.80 (0.03)
BucMuc	196	0.67(0.05)	0.55 (0.04)	0.78 (0.02)	0.77 (0.02)
HarPal	197	0.78(0.05)	0.68 (0.05)	0.80 (0.02)	0.78 (0.02)
LAntFos	51	0.58 (0.05)	0.57 (0.05)	0.68 (0.04)	0.66 (0.04)
LRetCre	123	0.64(0.07)	0.65 (0.05)	0.71 (0.03)	0.73 (0.04)
MidVag	45	0.61 (0.09)	0.58 (0.10)	0.69 (0.06)	0.71 (0.05)
PalTon	203	0.78 (0.03)	0.73 (0.03)	0.85 (0.01)	0.84 (0.02)
PosFor	22	0.60(0.15)	0.54 (0.10)	0.65 (0.07)	0.68 (0.09)
RAntFos	54	0.53(0.07)	0.52 (0.06)	0.64 (0.04)	0.59 (0.10)
RRetCre	85	0.56 (0.05)	0.57 (0.07)	0.67 (0.03)	0.67 (0.03)
Saliva	184	0.74 (0.03)	0.68 (0.03)	0.80 (0.02)	0.78 (0.01)
Stool	190	0.69 (0.06)	0.60 (0.07)	0.75 (0.03)	0.74 (0.03)
SubPla	205	0.78 (0.03)	0.72 (0.03)	0.87 (0.02)	0.84 (0.02)
SupPla	207	0.78 (0.04)	0.64 (0.04)	0.86 (0.02)	0.85 (0.02)
Throat	197	0.82 (0.03)	0.75 (0.03)	0.84 (0.02)	0.82 (0.02)
TonDor	207	0.81 (0.04)	0.65 (0.05)	0.88 (0.02)	0.86 (0.02)
VagInt	52	0.58 (0.07)	0.54 (0.07)	0.67 (0.05)	0.69 (0.06)

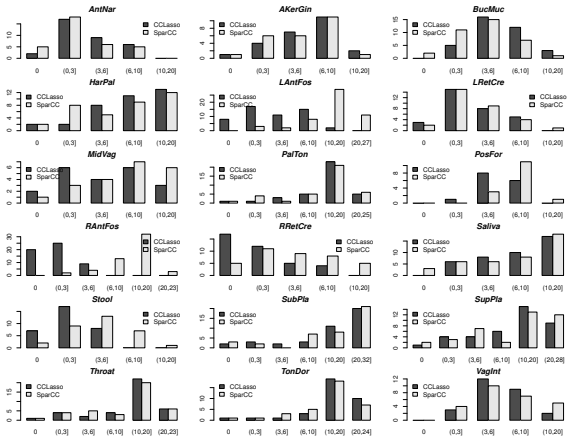

Fig. S3. Degree distribution of inferred correlation networks through CCLasso and SparCC (with a threshold 0.2).

Table S4. Pearson and Spearman correlations between CCLasso and SparCC for the HMP datasets.

Body Site	Sample Size	No. OTU	Pearson	Spearman
AntNar	152	34	0.93	0.93
AKerGin	193	25	0.96	0.96
BucMuc	196	36	0.96	0.97
HarPal	197	36	0.97	0.97
LAntFos	51	53	0.84	0.83
LRetCre	123	31	0.9	0.91
MidVag	45	21	0.9	0.92
PalTon	203	38	0.96	0.97
PosFor	22	15	0.9	0.89
RAntFos	54	54	0.71	0.66
RRetCre	85	38	0.85	0.85
Saliva	184	41	0.97	0.97
Stool	190	32	0.88	0.9
SubPla	205	41	0.98	0.98
SupPla	207	39	0.95	0.96
Throat	197	39	0.98	0.98
TonDor	207	35	0.96	0.96
VagInt	52	26	0.93	0.94

3.1.4 Correlation of estimated results between CCLasso and SparCC for HMP datasets Since consistent accuracy and reproducibility can only compare the algorithm sensitivity between CCLasso and SparCC for HMP datasets, we explore the relationship between these two methods from the whole estimated correlation matrix. Table S4 shows the Pearson and Spearman correlations of the estimation matrix between CCLasso and SparCC. We find all the correlations between CCLasso and SparCC are very high except some body sites such as right antecubital fossa.

3.2 Acid mine drainage dataset

It is interesting to explore the interaction network of microbes in some special environment. We use the proteome composition data using species-assigned protein counts from 28 microbial communities collected from an acid mine drainage environment (Mueller *et al.*, 2010). Figure S4 is the interaction network inferred from SparCC and CCLasso. We can find that SparCC has detected too many strong connections between unassigned groups and

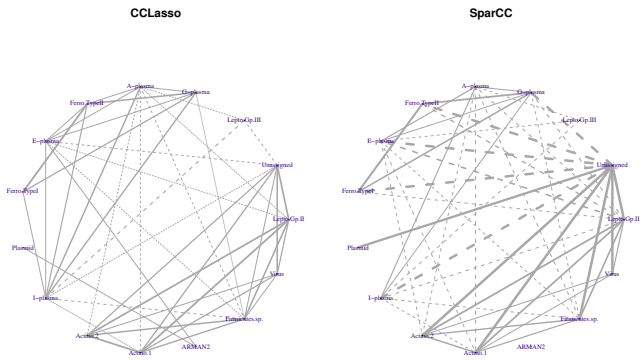

Fig. S4. Correlation network inferred from CCLasso and SparCC for acid mine drainage data. (The solid line means positive while dashed negative. And the wider the edge, the stronger the correlation. The correlation thresholds for CCLasso and SparCC are set 0.1 and 0.3 since the identification condition claims that at most 45 edges exist for only one possible sparse correlation matrix.)

others. Both CCLasso and SparCC have detected the triangle relationship among Ferro. Type I, Ferro. Type II and G-plasma.

REFERENCES

Mueller, R. S., Denef, V. J., Kalnejais, L. H., Suttle, K. B., Thomas, B. C., Wilmes, P., Smith, R. L., Nordstrom, D. K., McCleskey, R. B., Shah, M. B., *et al.* (2010). Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular systems biology*, **6**(374).