High-dimensional Linear State Space Models for Dynamic Microbial Interaction Networks

Iris Chen, Yogeshwar D. Kelkar, Yu Gu, Jie Zhou, Xing Qiu, Hulin Wu

October 11, 2019

Introduction

Microbial Interaction Network (MIN)

- Microbiome has direct relationship to human health. It is important to understand how bacteria that constitute the microbiom interact with their host and with each other.
- The complete set of interaction of various forms of co-operative and antagonistic relationships among bacteria can be depicted in the form of **Microbial Interaction Network (MIN)**
- The degree of relationships between two types of bacteria can be gauged from the impact that one type has over the growth and abundance of the other.

Previous Studies and Potential Problem

- Interactions among bacteria have traditionally been inferred using microbiological assays involving co-culturing.
  1. All bacteria can not be cultured.
  2. Laboratory inferred interactions may not occur in nature.
  3. Sequencing the variable regions of 16s ribosomal RNAs directly from biological samples gives estimates of abundance of large variety of bacteria.
- Few longitudinal studies have initially discovered the most abundant microbial taxa on various locations on human body.
  1. The construction of MIN operational in human microbiota remains a major challenge due to the high-dimensional and high-fluctuation nature of the data

Previous Studies and Potential Problem Contd.

- Many models have been proposed for constructing dynamic MIN, based on ordinary differential equation (ODE) models. An ODE model is formed by taking the derivative of bacterial abundance as a function of abundance of all other bacteria and/or external stimuli. This results in a directed network model, and the dynamic nature of MIN is automatically captured and quantified.
  1. It is computationally difficult to apply ODE models to many variables to simultaneously estimate system dynamics and regulatory relationships.
  2. most ODE models ignore both system and measurement errors, which in many cases have critical impact on results.

What They have Done?

- They focused on reconstructing dynamic MINs based on time series data.
- They develop a practical dynamic MIN reconstruction pipeline based on State Space Model (SSM) that not only incorporates many existing SSM parameter estimation and model selection techniques, but also is computationally efficient and applicable for "large p, small n" data
  1. A novel Expectation-Regularization-Maximization (ERM) computational framework for the SSMs is proposed, and they provide a feasible implementation strategy for initialization of the ERM algorithm.
  2. The vectorization of the matrices in the SSM has been done and a concept of "pseudo-regression" is used to justify the R step for L1-regularization based on which the standard LARS algorithm with minor modifications can be carried out.
  3. A new row-based algorithm is proposed in order to reduce the memory footprint, which is a major computational cost for high-dimensional data analysis.

Methods

State Space Model (SSM)

- SSM is a special case of Dynamic Bayesian Networks (DBN).
- Here only linear SSM is considered which is also known as Linear Dynamic System (LDS).
- Let $y_t \in \mathbb{R}^p$ represent a p-dimensional vector of microbial abundance of p bacterial operational taxonomic units (OTUs) observed at time t.
- y_t is assumed to be generated from a k-dimensional real-valued hidden state variable vector $x_t \in \mathbb{R}^k$,and the sequence of evolving x_t follows a first-order Markov process.

Methods

SSM Contd..

- For $t = 1, \cdots, T$ the model is as follows:

$$x_t = Ax_{t-1} + w_t$$
$$y_t = Cx_t + v_t$$

- Here A is the $k \times k$-dimensional state dynamic (AR) matrix C is the $p \times k$-dimensional observation matrix and $w_t \sim N(0, Q), v_t \sim N(0, R)$ are independent system and measurement noises, respectively. Both Q and R are assumed to be diagonal in many practical applications. The initial state vector $x_0 \in \mathbb{R}^k$ is usually assumed to have distribution $N(\mu, \Sigma)$.
- In our case, $c = I_{p\times p}, k = p, Q = \sigma_Q^2 I_{p\times p}, R = \sigma_R^2 I_{p\times p}$
- Each element a_{ij} in $p \times p$ system matrix A represents a directed edge in the network which is time-invariant, and reflects the interacting effect from bacterial species j to bacterial species i.

Methods

Joint Likelihood

- The joint likelihood for complete data for the SSM is

$$P(\theta) = P(x_1) \prod_{t=2}^{T} P(x_t|x_{t-1}) \prod_{t=1}^{T} P(y_t|x_t)$$

- $\theta = (A, Q, R, \mu, \Sigma)$
- The joint log-likelihood is

$$\begin{aligned}\log P(\theta) = &- \sum_{t=1}^{T} \frac{1}{2}(y_t - x_t)' R^{-1}(y_t - x_t) - \frac{T}{2} \log |R| \\ &- \sum_{t=2}^{T} \frac{1}{2}(x_t - Ax_{t-1})' Q^{-1}(x_t - Ax_{t-1}) - \frac{T-1}{2} \log |Q| \\ &- \frac{1}{2}(x_1 - \mu)' \Sigma^{-1}(x_1 - \mu) - \frac{1}{2} \log |\Sigma| - Tp \log(2\pi)\end{aligned}$$

Methods

ERM Method

The Expectation-Regularization-Maximization (ERM) algorithm is used to estimate the high-dimensional sparse system matrix A, and other parameters, as well as the state variables for linear SSMs. The procedure is outlined as follows:

- E Step:The conditional expectation of the likelihood is calculated by

$$G(\theta|\theta^{(r-1)}) = E_{x|Y,\theta^{(r-1)}} \log P(\theta)$$

- R Step:Adaptive LASSO method, is employed to obtain the estimate of the sparse system matrix A denoted by $A^{(r)}$.
- M Step:The MLE of other model parameters $\theta^{\star} = (Q, R, \mu, \Sigma)$ denoted by $\theta^{\star(r)}$ is obtained by maximizing the conditional expectation of likelihood

$$G(\theta^{\star}|\theta^{(r-1)}, A^{(r)}) = E_{x|Y,\theta^{(r-1)},A^{(r)}} \log P(\theta)$$

- Notation:

$$
\begin{aligned}
x_t^T &= E(x_t|y_1, \cdots, y_T) \\
(xx')_t^T &= E(x_t x_t'|y_1, \cdots, y_T) \\
(xx')_{t,t-1}^T &= E(X_t x_{t-1}'|y_1, \cdots, y_T) \\
V_t^\tau &= Var(x_t|y_1, \cdots, y_\tau) \\
V_{t,t-1}^\tau &= Cov(x_t x_{t-1}'|y_1, \cdots, y_\tau)
\end{aligned}
$$

- τ is an arbitrary time point.

$$
\begin{aligned}
x_t^{t-1} &= A x_{t-1}^{t-1} \\
V_t^{t-1} &= A V_{t-1}^{t-1} A' + Q \\
K_t &= V_t^{t-1} C' (V_t^{t-1} C' + R)^{-1} \\
x_t^t &= x_t^{t-1} + K_t (y_t - C x_t^{t-1}) \\
V_t^t &= V_t^{t-1} - K_t C V_t^{t-1}
\end{aligned}
$$

In the first iteration A will be estimated by an initial R step. The initial state mean $x_0^0 = \mu$ can be replaced by a small non-zero initial vector and the variance matrix $V_0^0 = \Sigma$ can be replaced by a small diagonal matrix. R, Q can be initialized by two identity matrices.

$$
\begin{aligned}
J_{t-1} &= V_{t-1}^{t-1} A' (V_t^{t-1})^{-1} \\
x_{t-1}^T &= x_{t-1}^{t-1} + J_{t-1}(x_t^T - A x_{t-1}^{t-1}) \\
V_{t-1}^T &= V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1}) J_{t-1}' \\
V_{t-1,t-2}^T &= V_{t-1}^{t-1} J_{t-2}' + J_{t-1}(V_{t,t-1}^T - A V_{t-1}^{t-1}) J_{t-2}'
\end{aligned}
$$

where $V_{t-1,t-2}^T$ is initialized by $V_{T,T-1}^T = (I - K_T C) A V_{t-1}^{t-1}$. Alternatively, the matrix V_t^t in forward recursion and the matrix $(V_t^{(t-1)})^{-1}$ in backward recursion can be expressed as

$$
\begin{aligned}
V_t^t &= [C' R^{-1} C + (V_t^{t-1})^{-1}]^{-1} \\
(V_t^{t-1})^{-1} &= Q^{-1} - Q^{-1} A [4 A' Q^{-1} A + (V_{t-1}^{t-1})^{-1}]^{-1} A' Q^{-1}
\end{aligned}
$$

ERM Method

R Step

- Adaptive LASSO estimates can be implemented by using the LARS Algorithm.
- In order to modify and apply the LARS algorithm for the R step in the algorithm the following notations are used:

$$X^\star = (x_2, x_3, \cdots, x_T), \qquad X = vec(X^\star)$$
$$Z^\star = (x_1, x_2, \cdots, x_{t-1})', \qquad Z = Z^\star \otimes I_{p\times p}$$
$$a^\star = (a_1, a_2, \cdots, a_p), \qquad \alpha = vec(a^\star)$$
$$e^\star = (w_2, w_3, \cdots, w_T), \qquad e = vec(e^\star)$$

- a_i is the ith row vector of A, α is the vectorized A which is a $(p^2 \times 1)$ vector, e is a $(p(t-1)\times 1)$ vectors that represent measurement errors.

Matrix-based ERM Algorithm

- We want to minimize

$$G(\alpha) = E_{X,Z|Y,\theta^{(r-1)}}\{(X - Z\alpha)'(X - Z\alpha)\}$$

- The MLE of A is

$$\hat{A} = \left(\sum_{t=2}^{T}(xx')_{t,t-1}^{T}\right)\left(\sum_{t=1}^{T-1}(xx')_{t}^{T}\right)^{-1}$$

- An $L1$-regularized estimator of A that minimizes

$$G(\alpha) = E_{X,Z|Y,\theta^{(r-1)}}\{(X - Z\alpha)'(X - Z\alpha)\} + \lambda\sum_{j}\hat{w}_j|\alpha_j|$$

where λ is a tuning parameter.

Matrix-based ERM Algorithm

Choosing Weight and Tuning Parameter

- If $T-1 > p$, the MLE of A is

$$\hat{w}_{ij} = |\hat{A}_{ij}|^{-1}$$

- If $T-1 \leq p$,

$$\hat{w}_{ij} = |\tilde{A}_{ij}|^{-1}$$

$$\tilde{A} = \left(\sum_{t=2}^{T} (xx')_{t,t-1}^{T} \right) diag \left(\sum_{t=1}^{T-1} (xx')_{t}^{T} \right)^{-1}$$

- The elements of $\hat{A}$ are shrunk towards 0 as λ increases.
- The extended BIC is recommended since it contains an extra penalty term with the consideration of different prior distributions over the model space.

Matrix-based ERM Algorithm

M-STep

The estimates for the remaining parameters are

$$\hat{\mu} = x_1^T$$

$$\hat{\Sigma} = (xx')_1^T - x_1^T x_1^{T'}$$

$$\hat{R} = \frac{1}{T}\sum_{t=1}^{T}(y_t y_t' - x_t^T y_t')$$

$$\hat{Q} = \frac{1}{T-1}\sum_{t=2}^{T}\left((xx')_t^T - \hat{A}_{aL}(xx')_{t,t-1}^T\right)$$

Row-based ERM Algorithm

- Let X_i as the ith row of $X^\star$, a_i as the ith row of A and e_i as the ith row of $e^\star$.
- The adaptive Lasso can be similarly used using

$$X_i = Z^\star a_i + e_i$$

where $Z^\star = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(T-1)1} & x_{(T-1)2} & \cdots & x_{(T-1)p} \end{pmatrix}$

- The adaptive LASSO estimate of $a_i (i = 1, \cdots, p)$ is

$$\hat{a}_{iaL} = \arg\min_{a_i} E_{X,Z|Y,\theta^{(r-1)}} \{ ||X_i - Z^\star a_i||^2 + \lambda \sum_{ij} \hat{w}_{ij|a_{ij}} \}$$

Results

Simulation Studies

- The performance of the proposed ERM algorithm for variable selection, False Positive Rate (FP) and False Negative Rate (FN) of $\hat{A}_{aL}$ by

$$FP = \frac{\sum_{ij} I_{\{\hat{a}_{ij} \neq 0 | a_{ij} = 0\}}(\hat{a}_{ij})}{N}$$

$$FN = \frac{\sum_{ij} I_{\{\hat{a}_{ij} = 0 | a_{ij} \neq 0\}}(\hat{a}_{ij})}{P}$$

- P is the number of nonzero elements and N is the number of zero elements in A.
- The nonzero elements of A were generated form $\overset{+}{-}(0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$.
- $Q = I$ and $R = 0.1I$ for all the cases and the number of time points $T = 60$.
- They have started the ERM Algorithm from the R step to the 100 simulated data sets.

Simulation Studies

p	p^2	% nonzero	algorithm	FP	FN
8	64	23.44	row	0.0459	0.0600
			matrix	0.0255	0.1073
20	400	8.75	row	0.0108	0.0531
			matrix	0.0095	0.0714
50	2500	3.36	row	0.0053	0.1458
			matrix	0.0039	0.1760
80	6400	2.34	row	0.0060	0.2306
			matrix	*N/A*	*N/A*

https://doi.org/10.1371/journal.pone.0187822.t001

Table 1. Simulation results: Comparisons of variable selection performance between the row-based and matrix-based ERM algorithms

Simulation Studies

T	Q	R	FP	FN
20	fixed as true	fixed as true	0.0167	0.3111
	estimated	fixed as true	0.0192	0.3479
50	fixed as true	fixed as true	0.0151	0.1617
	estimated	fixed as true	0.0150	0.1803
100	fixed as true	fixed as true	0.0189	0.0915
	estimated	fixed as true	0.0180	0.0984

https://doi.org/10.1371/journal.pone.0187822.t002

Table 2. Evaluation of the row-based ERM algorithm for variable selection with respect to number of time points T and Q estimation. $p = 41$.

Applications to Microbiota Data

- They have applied their proposed method to investigate the dynamic interactions among vaginal bacteria.
- The magnitude for abundance varies widely for different bacteria.
- Without proper standardization, a uniform L1 penalty is much more likely to set the edges related to less abundant OTUs to 0 and results in a simplistic networks dominated by a few most abundant OTUs.
- To overcome the problem they have proposed to standardize the data

$$Y_{ij} = \frac{\tilde{Y}_{ij} - \bar{Y}_i}{sd(\tilde{Y}_i)}$$

Results

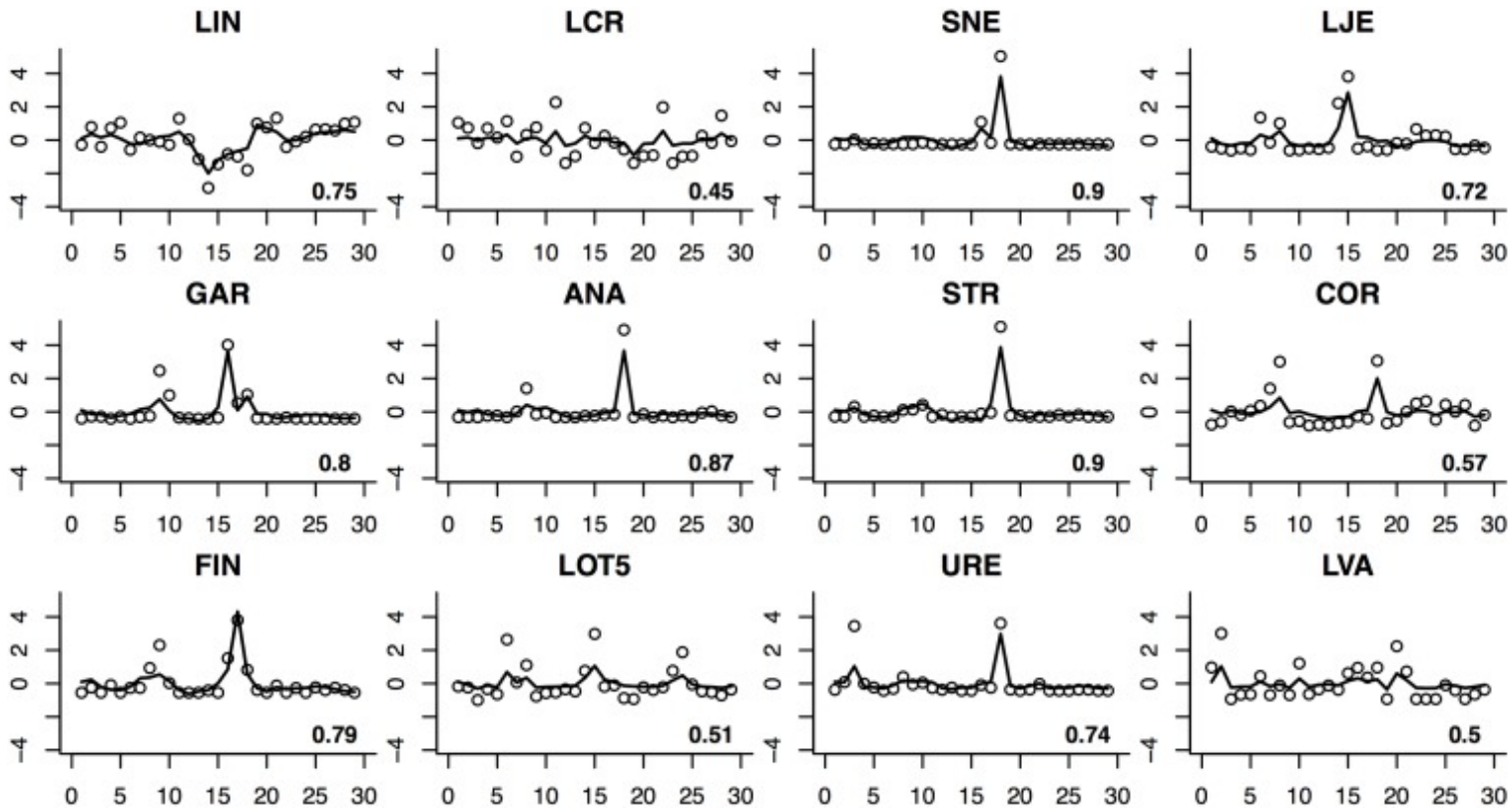

Fig 1. One-step-ahead prediction for subject 15. Each cell represents prediction for a different OTU. Solid lines depict the predicted values whereas circles indicate standardized temporal abundances of OTUs. Abbreviations for OTUs: LIN Lactobacillus iners; LCR Lactobacillus crispatus; SNE Sneathia sp.; LJE Lactobacillus jensenii; GAR Gardnerella sp.; ANA Anaerococcus sp.; STR Streptococcus sp.; COR Corynebacterium sp.; FIN Finegoldia sp.; LOT5 Lactobacillus otu5; URE Ureaplasma sp.; LVA Lactobacillus vaginalis.

Results

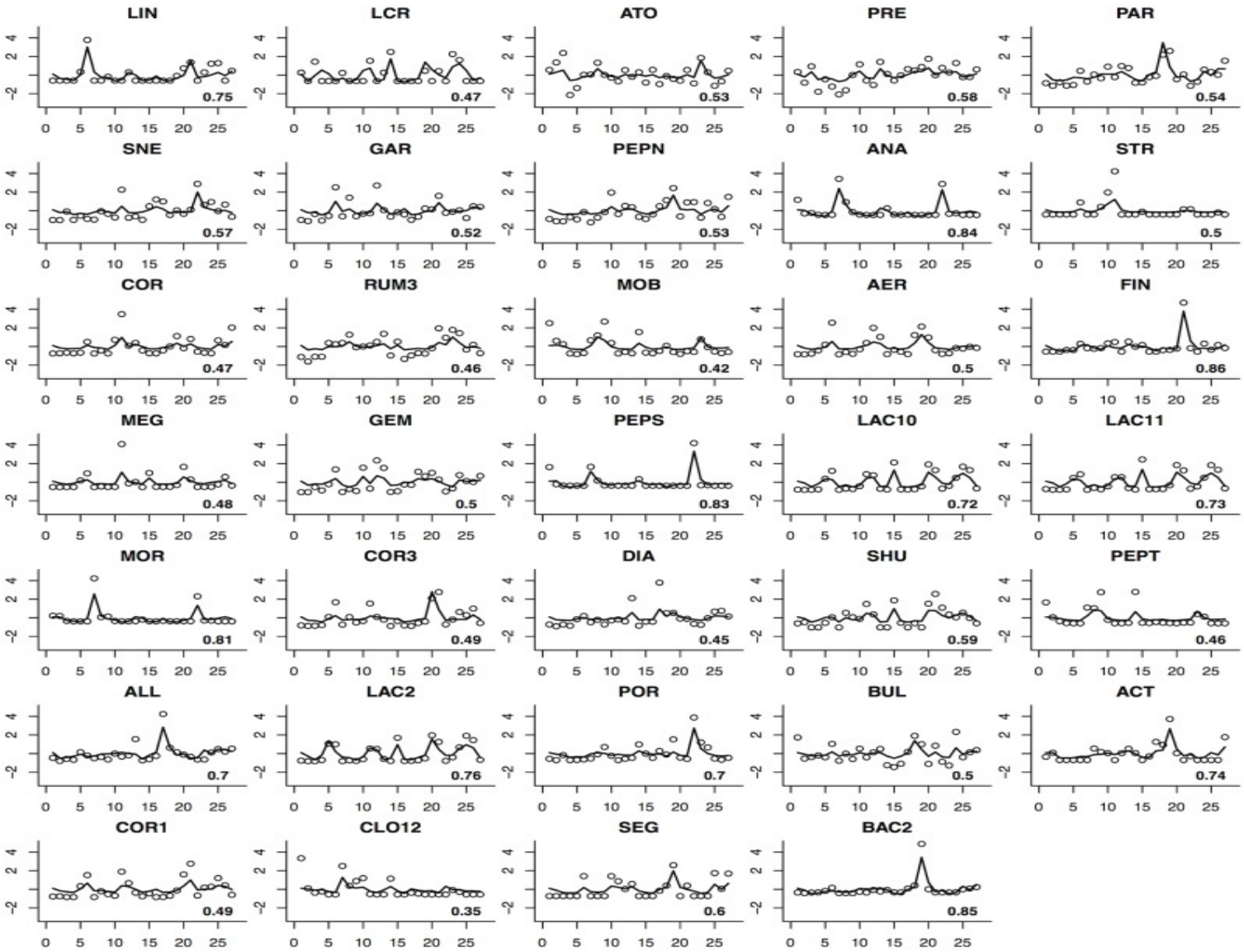

Fig 2. One-step-ahead prediction for subject 6. Each cell represents prediction for a different OTU. Solid lines depict the predicted values whereas circles indicate standardized temporal abundances of OTUs. Abbreviations for OTUs: LIN

Results

Bacteria	Positive Effects	Negative Effects
LIN	LIN	LJE, FIN, LOT5
LCR	FIN	
SNE	FIN	LCR
LJE	SNE, GAR, FIN	STR
GAR	FIN	SNE, ANA, STR, URE
ANA		SNE, ANA, STR, FIN, URE
STR	LIN	FIN
COR	FIN	SNE, LVA
FIN	SNE, GAR, ANA, STR, COR, FIN, URE	
LOT5	FIN	
URE		FIN
LVA	FIN, URE	

https://doi.org/10.1371/journal.pone.0187822.t003

Table 3. Interactions among bacteria for subject 15. See legend for Fig 1 for OTU abbreviations.

Results

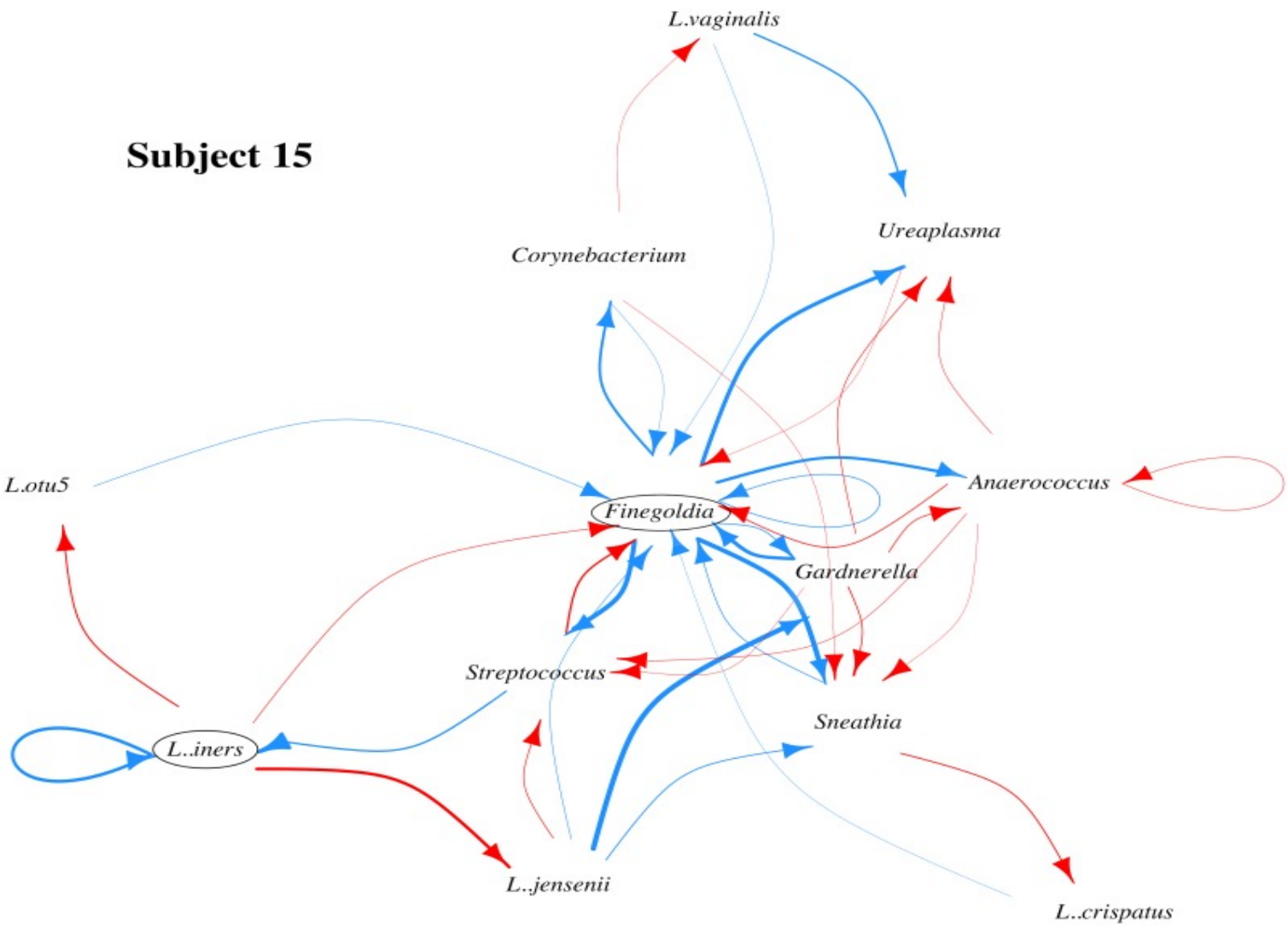

Fig 3. Microbial interaction network (MIN) for subject 15. Blue and red arrows indicate directed positive and negative effects respectively. Arrow width indicates effect magnitude. Circles highlight bacterial species that impact multiple other species in the MIN and whose critical role in the MIN has either experimental support in literature (L. iners) or has never been recognized before (Finegoldia sp.).

Results

Bacteria	Positive Effects	Negative Effects
LIN	ANA, PEPS, MOR, CLO12	
LCR	MEG, LAC10, LAC11, SHU, LAC2	
PRE	LCR	
PAR	PEPN, AER, DIA, ACT, SEG, BAC2	
SNE	ALL	
ANA	MOB, PEPT	GEM
STR	STR	
FIN	SNE, ANA, PEPS, POR	
GEM	COR	
PEPS	ATO, RUM3	
COR3	FIN, SHU	
ALL	PAR, BUL	
LAC2	LIN, GAR, COR1	CLO12
ACT	PRE	
BAC2	COR3	

https://doi.org/10.1371/journal.pone.0187822.t004

Table 4. Interactions among bacteria for subject 6. See legend for Fig 2 for OTU abbreviations

Results

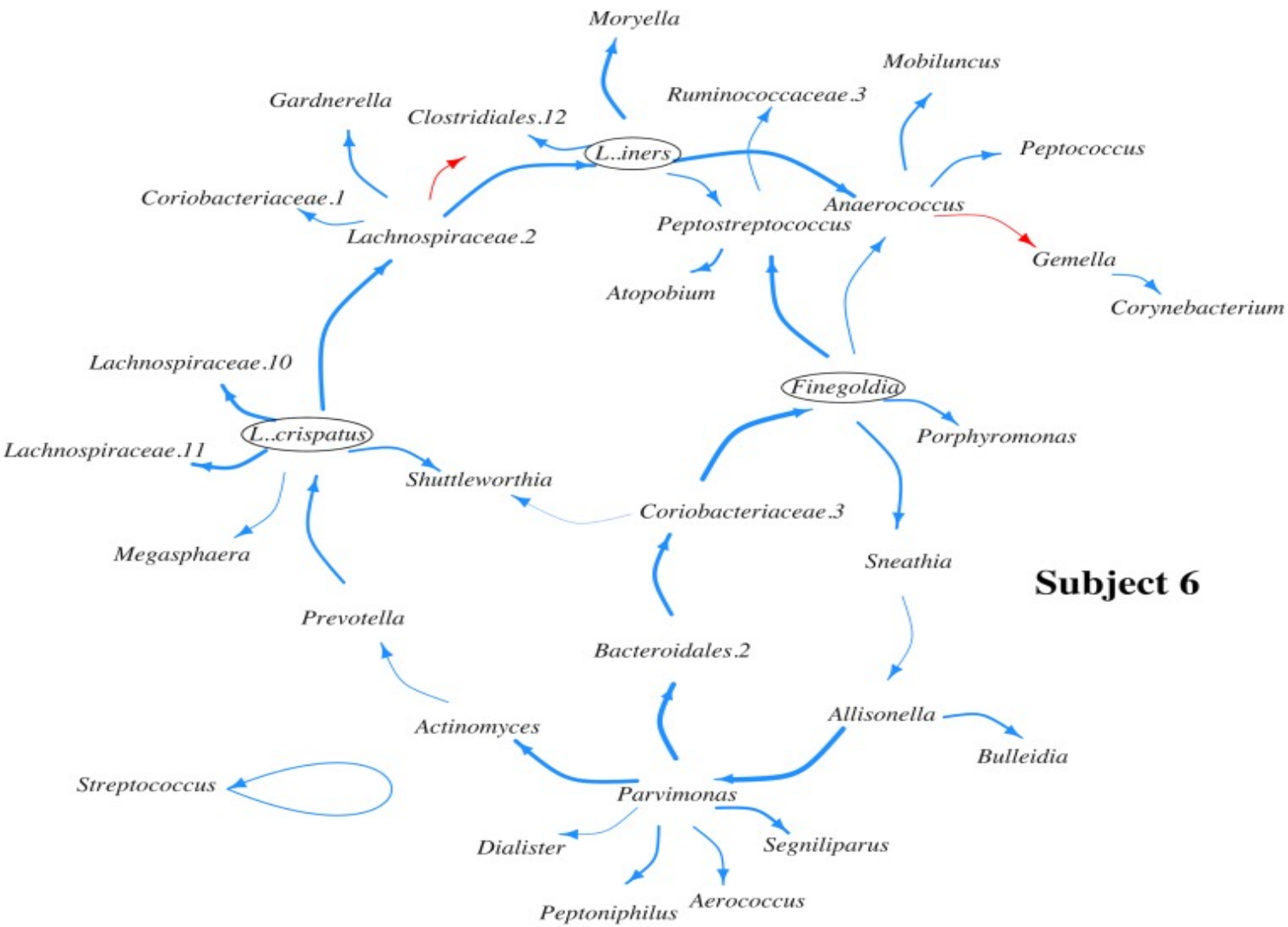

Fig 4. Microbial interaction network (MIN) for subject 6. Blue and red arrows indicate directed positive and negative effects respectively. Arrow width indicates effect magnitude. Circles highlight bacterial species that impact multiple other species in the MIN and whose critical role in the MIN has either experimental support in literature (L. iners) or has never been recognized before (Finegoldia sp. and L. crispatus)

Summary

Pros

- A new ERM algorithm is proposed for a high dimensional linear state space model to construct dynamic MINs from time course microbiome data.
- To overcome the difficulty for large matrix manipulations for the high-dimensional SSMs, a row-based ERM algorithm was proposed and evaluated against the standard matrix-based algorithm.
- The method works well for both simulated and real life data.

Summary

Weaknesses

- The state-space model considers time as discrete steps instead of a continuous variable as used in many alternative network models.
- The model should only be applied to modeling the interactions between key OTUs with low sparsity.
- In this paper they considered estimating MINs for each individual. It is more meaningful to construct the common MIN for a population which will require advanced SSM.
- Addition of the R step may change the theoretical properties which needs further investigations.