The Challenges of Estimating the Interactions between Bacterial Species from Longitudinal Data

Thierry Van Effelterre – Luc Bijnens
Janssen R & D

NCS 2018
Paris

The microbiome is a complex ecosystem

- The microbiome is made up of millions of bacteria and thousands of different bacterial species,

  although different species might be unequally present and abundant.

- The abundance of a bacterial species depends on
  - its own growth
  - how it interacts with the other bacterial species

    … at least some of them.

It is a **complex ecosystem**

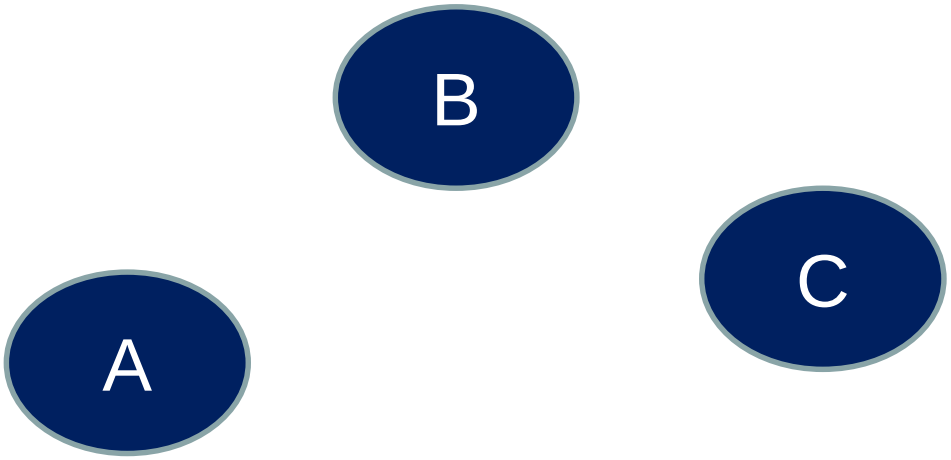
→ should better be studied as such.

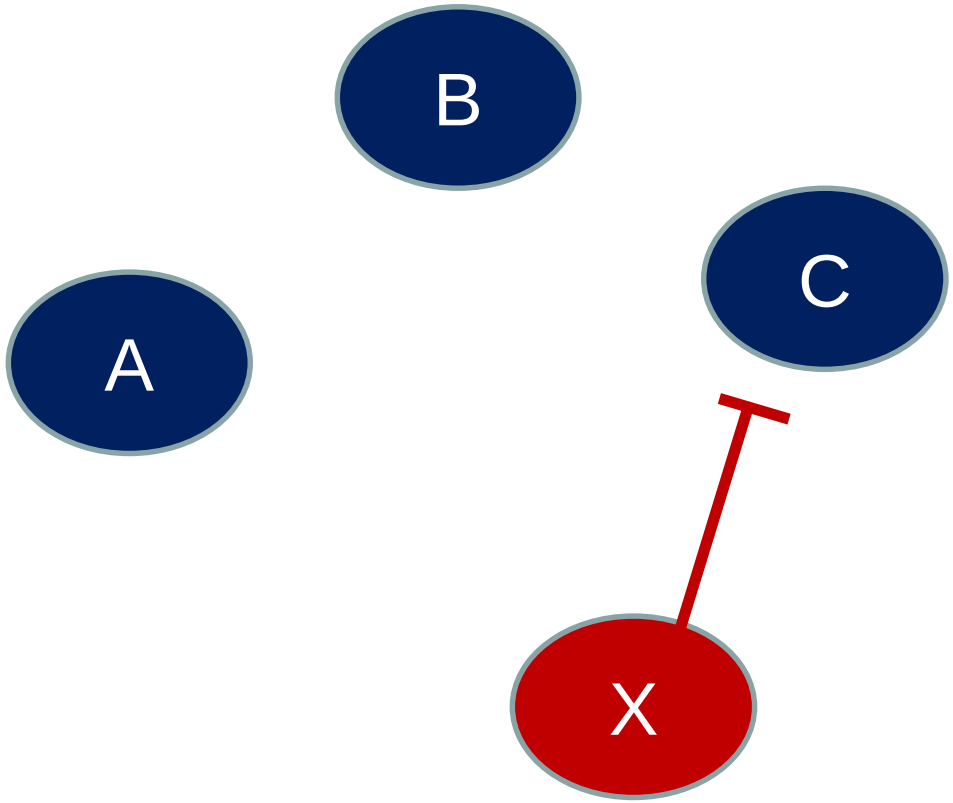
Microbiome-targeted interventions

- Probiotics and life therapeutics, targeting the microbiome to improve health, will affect
  - the abundances of « targeted » bacterial species
  - also, potentially, other bacterial species interacting with them.

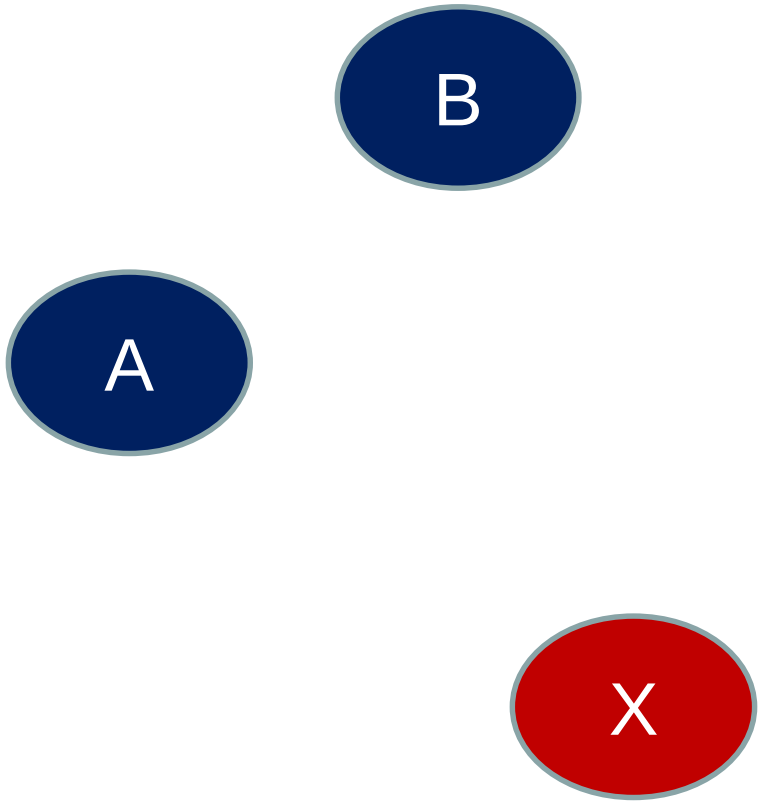
→ The dynamics and longer-term evolution of the abundances of the different species, in particular whether they persist or are eliminated, may be quite complex.

Time of probiotics administration

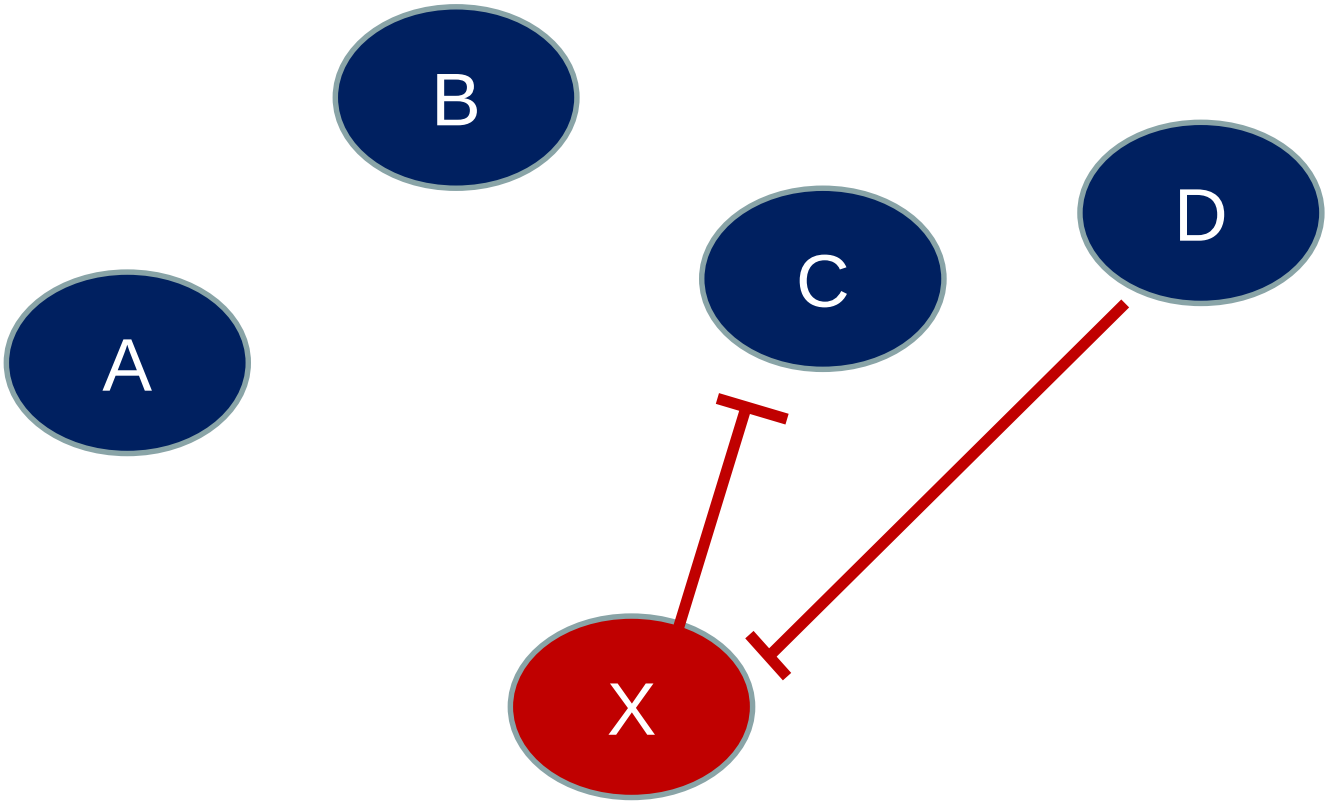



B
C
A
X

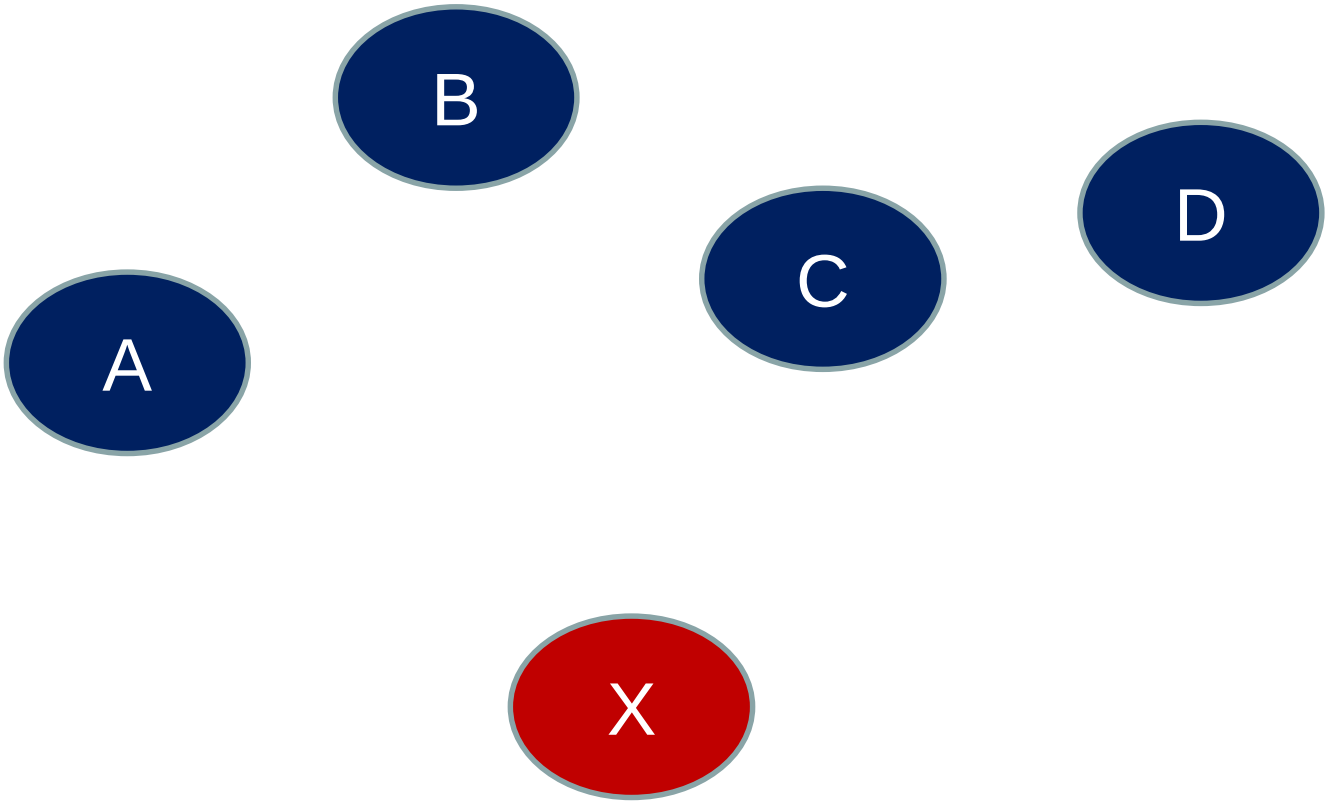
… some time after probiotics administration …



Time of probiotics administration



… some time after probiotics administration …



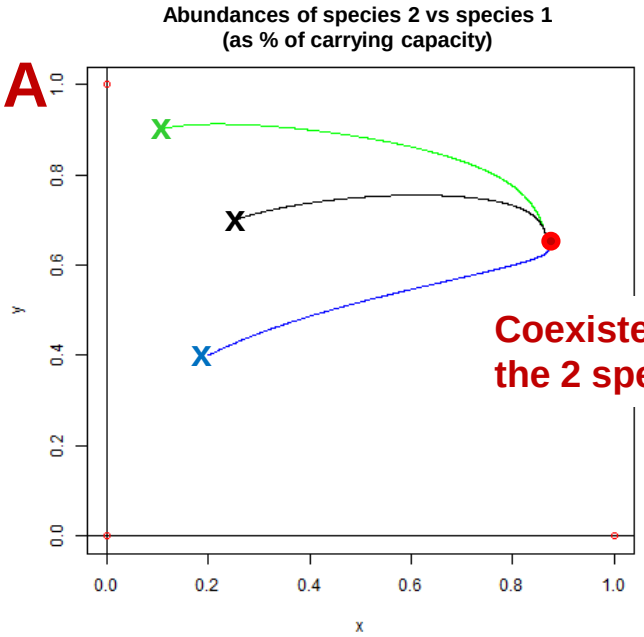
Dynamical systems

- A qualitative and quantitative understanding of a complex ecosystem is the realm of dynamical systems.

- This framework allows to model the time evolution of the abundances of different bacterial species in a mechanistic way.

- In particular, it allows to better understand
  - towards which steady state the abundances of the different species will evolve in the long term,
  - whether a species will persist or be eliminated.

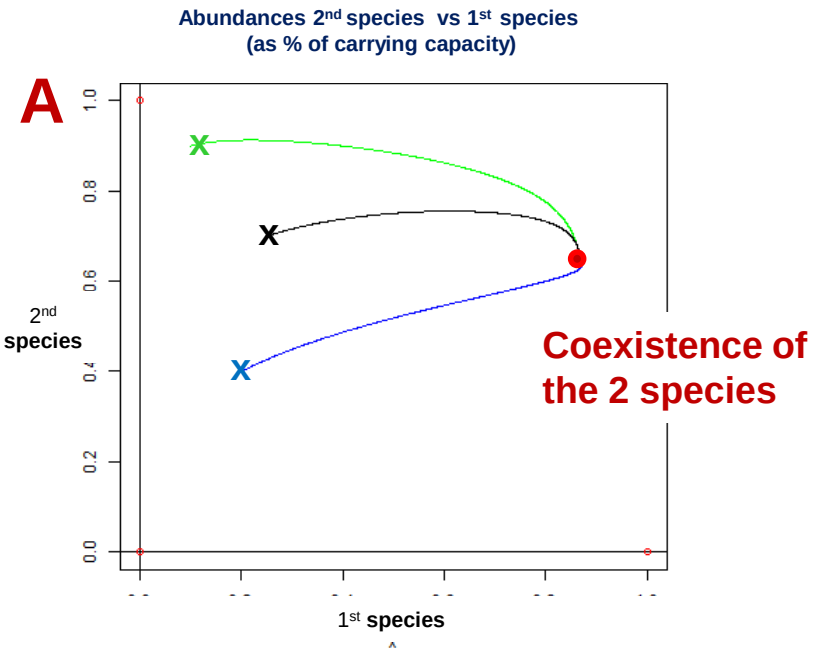
Competitive Lotka-Volterra with 2 species



Coexistence of the 2 species

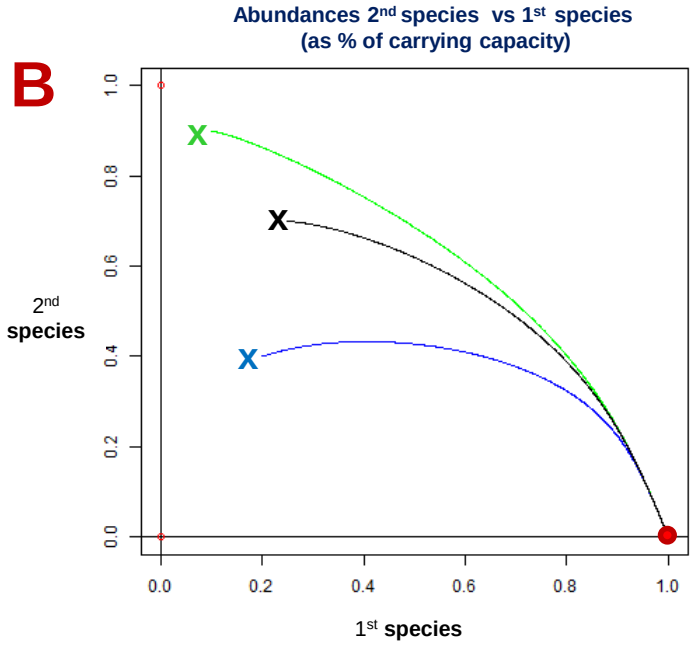
Competitive Lotka-Volterra with 2 species

A



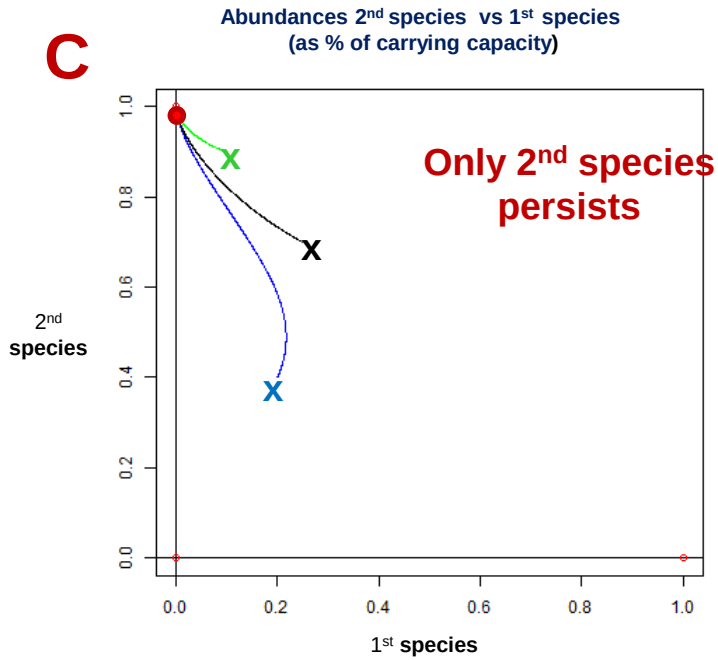
Coexistence of the 2 species

B



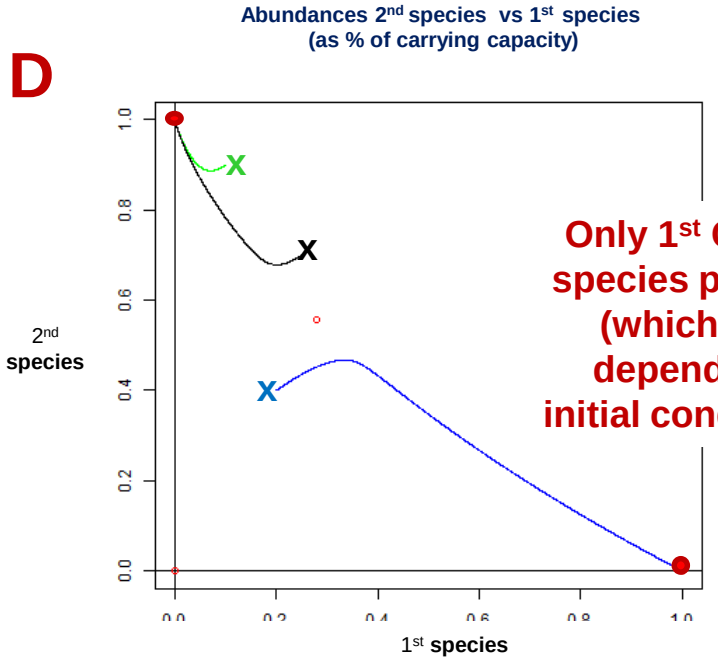
Only 1st species persists

C



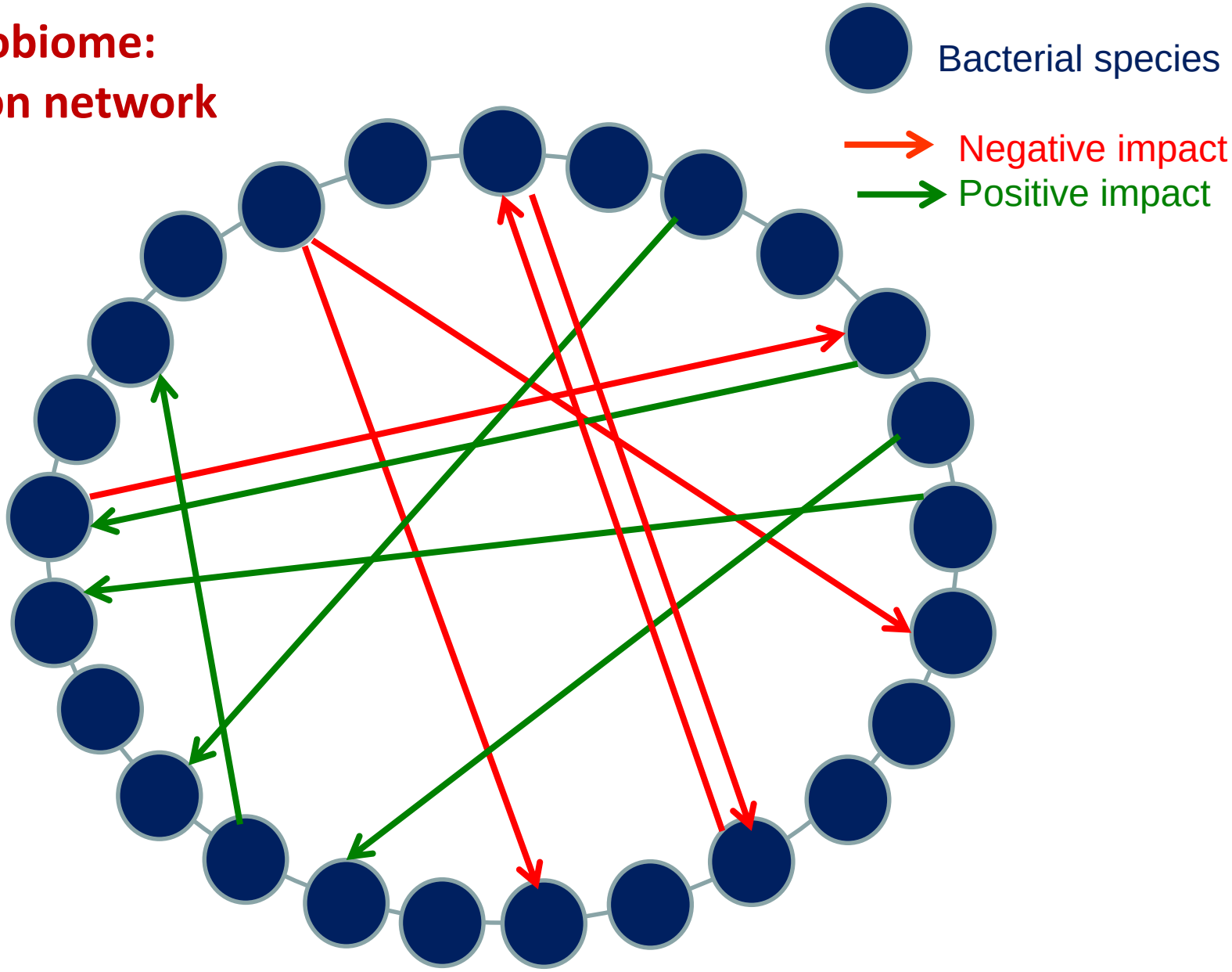
Only 2nd species persists

D



Only 1st OR 2nd species persists (which one depends on initial conditions)

The microbiome: Interaction network



How can we gain information about the interactions between species?

- Interventions targeting the microbiome may dynamically affect the whole community of bacteria.

- Therefore of value to better understand and quantify the growth rates and strengths (and sign) of the interactions between bacterial species.

- Because this is a dynamic process, it was proposed by several authors to estimate the growth rates and interactions from longitudinal data.

This is the case for example from 16S RNA data from stool samples collected over time.

The Generalized Lotka-Volterra Model

For a system of N "species", with **time-varying abundances** $P_1(t), P_2(t), \ldots, P_N(t)$

The **Generalized Lotka-Volterra (GLV) model** is characterized by the following system of N ordinary differential equations (ODE's):

$$\frac{dP_i(t)}{dt} = P_i \left(g_i + \sum_{j=1}^{N} \alpha_{ij} P_j\right) \qquad for \quad i = 1, 2, \ldots, N$$

Where

$g_1, g_2, \ldots, g_N$ are the intrisinc growth rates of the N species $\quad i = 1,2,\ldots, N$

α_{ij} is the strength of the interaction between species *i* and species *j*: more specifically how the abundance of species *j* affects the growth rate of species *i*

$$i = 1,2,\ldots, N;\ j = 1,2,\ldots, N \qquad i \neq j$$

α_{ii} is the strength of the self-interaction for species *i*, with $\alpha_{ii} < 0$

Methods based on the discretization of the GLV system

Given observations over time for the abundance of each species, e.g.

$$P_1(t_1), P_1(t_2), \ldots, P_1(t_{N1}), P_2(t_1), P_2(t_2), P_2(t_{N2}), \ldots, P_N(t_1), P_N(t_2), P_N(t_{NN})$$

This system can be discretized and reduces to a system linear in the N state variables.

$$\frac{Ln(P_i(t+\Delta t)) - Ln(P_i(t))}{\Delta t} = g_i + \sum_{j=1}^{N} \alpha_{ij} P_j(t)$$

Different methods can be used to estimate the model parameters from this linearized system:

- Ridge regression [Stein et al, PLOS Comp. Biology 9(12), 2013
  Bucci et al, Genome Biology 17:121, 2016]
- Bayesian methods [Bucci et al, Genome Biology 17:121, 2016]
- Stepwise regression with "bagging" [Fisher & Mehta, PLoS One, 9(7), 2014]

The approach we propose here ...
uses the linearity of the GLV model in its model parameters

$$\frac{dP_1(t)}{dt} = P_1 \left(g_1 + \sum_{j=1}^{N} \alpha_{1j} P_j\right)$$

$$\frac{dP_2(t)}{dt} = P_2 \left(g_2 + \sum_{j=1}^{N} \alpha_{2j} P_j\right)$$

...

$$\frac{dP_N(t)}{dt} = P_N \left(g_N + \sum_{j=1}^{N} \alpha_{Nj} P_j\right)$$

Although non-linear in the state variables, this system of ODE's **is linear in the model parameters**

→Can use a estimation approach by [Dattner & Klaassen, Electronic Journal of Statistics, Vol. 9, 2015]

"Smooth ans integrate" method to estimate the model parameters

$$\frac{d\vec{P}(t)}{dt} = g(\vec{P}(t))\ \vec{\vartheta} \quad \Rightarrow \vec{P}(t) = \vec{\zeta} + \int_0^t g(\vec{P}(s))ds\ \vec{\vartheta}$$

For t in [0, 1]

where $\vec{\zeta} = (\zeta_1, \zeta_2, ..., \zeta_N)$ is the vector with the initial conditions for the N species

$Let\ \hat{P}\ (t)$ **be an estimator of P(t) based on the observations**

→ **Estimate the parameters and the initial conditions by minimizing**

$$\int_0^1 \| \hat{P}\ (t) - \zeta - \int_0^t g(\hat{P}(s))ds\ \eta \|^2\ dt$$

with respecto to η **and** ς

From: Dattner & Klaassen, Electronic Journal of Statistics, Vol. 9, 2015

"Smooth and Integrate" approach

In general, we may measure abundances for multiple (say *B*) individuals (or animals). We model the abundances of the *N* bacterial species in the *B* individuals.

We use the method **species by species** because the *N* + 1 parameters affecting a given species are decoupled from those affecting another species in the GLV model.

→ For each species:

STEP 1: For each individual *j*

smooth the observed data points $O_{i,j,k}$ **using local polynomials**

→ Estimators for the population abundance $\hat{P}_{ij}(t)$

STEP2: **Use those estimator to estimate the model parameters and the initial conditions (by individual) for species *i* by direct (numerical integration).**

Simulations

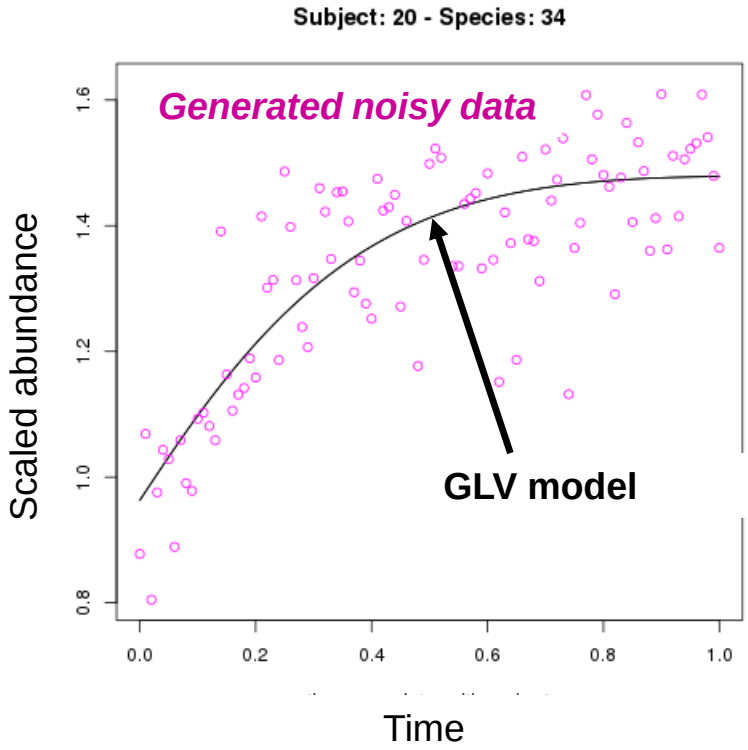
We evaluated the method on simulated data:

The Steps:

1. Generate model parameters (growth and interactions between species)

2. Solve the dynamic (ODE) system of the GLV model over time with those parameters for a group of subjects, each subject with own initial conditions.

3. Add noise (assumed additive with normal distribution) to abundances at determined number of equally spaced time points along the GLV model trajectory.



4. Estimate the GLV model parameters with the "smooth and integrate method" using the generated noisy longitudinal data.

Simulations

We first assessed **the influence of a few key "design parameters"** on the quality of estimation, using the "smooth and integrate method":

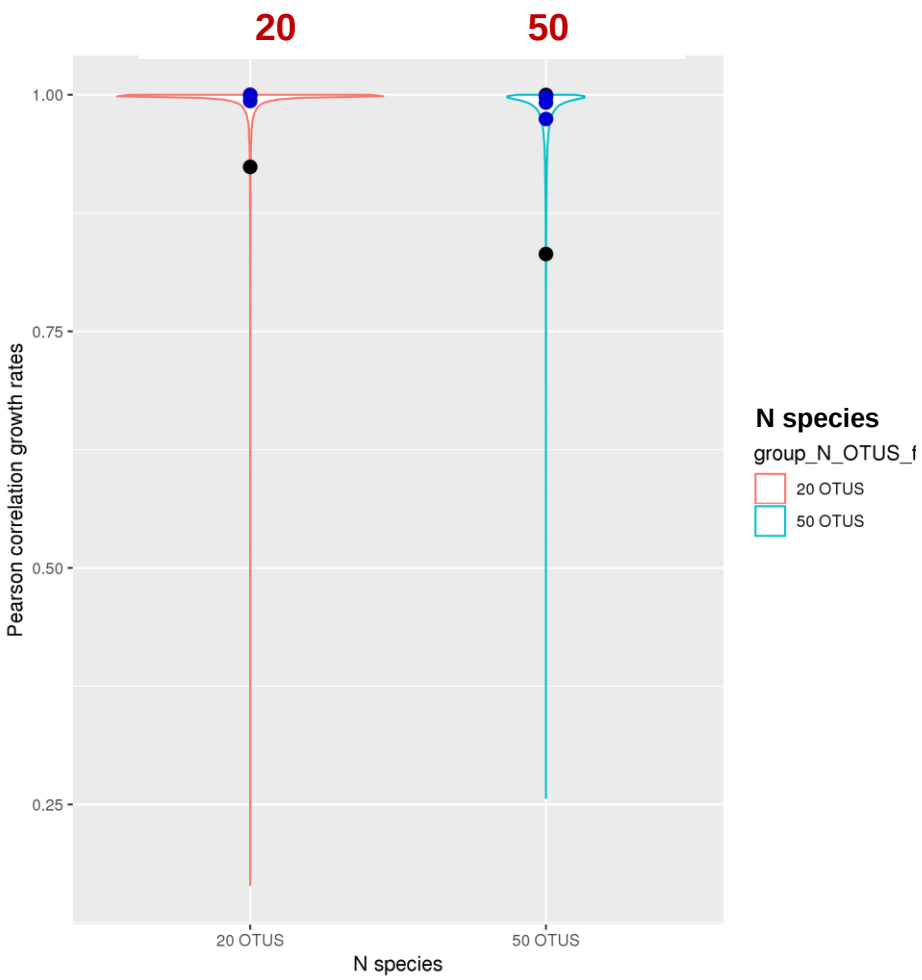
- Total number of species considered
- Number of subjects
- Number of samples ("points") per subject over time
- Magnitute of noise of the simulated data.

For each "scenario", estimated and true parameters were compared with:

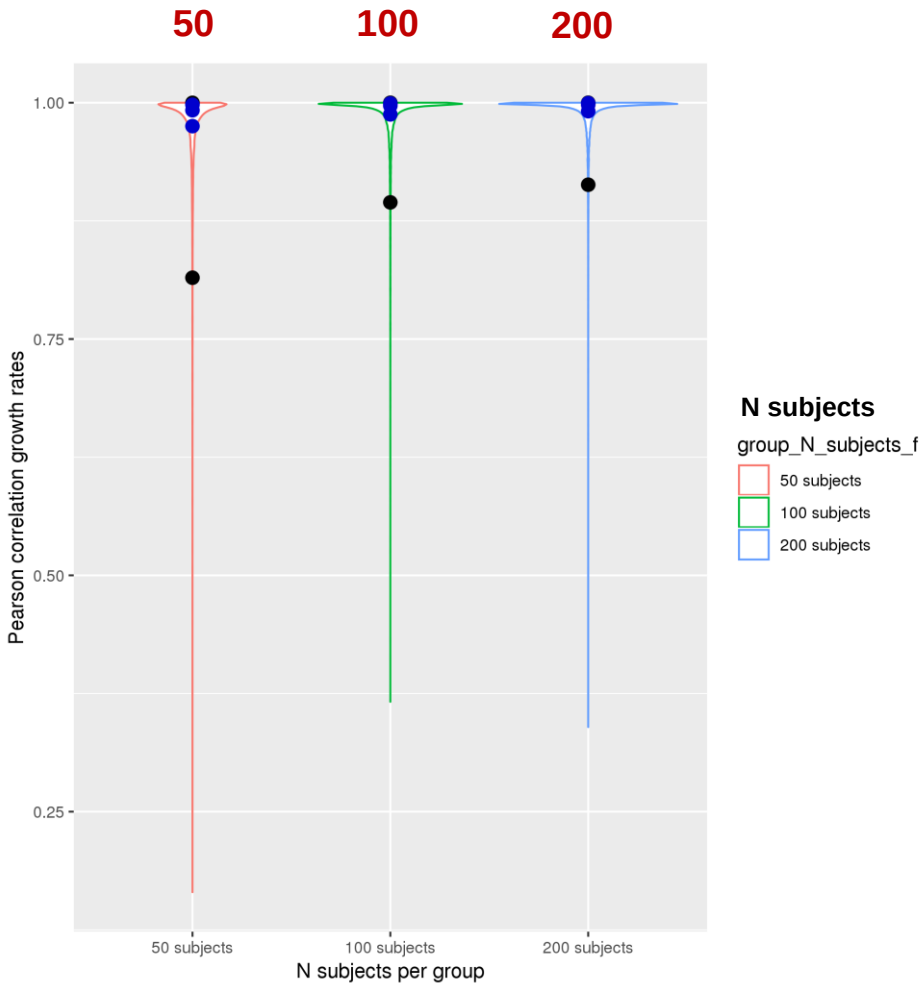
- Correlations between estimated and original parameters
  - growth parameters
  - pairwise interactions between species
- Comparison of signs of pairwise interactions:
  - percentage of pairwise interactions with the same sign (-, 0 or +)
  - Cohen's Kappa statistic for agreement between signs

Pearson correlation between estimated and original **growth rates**

By number of species



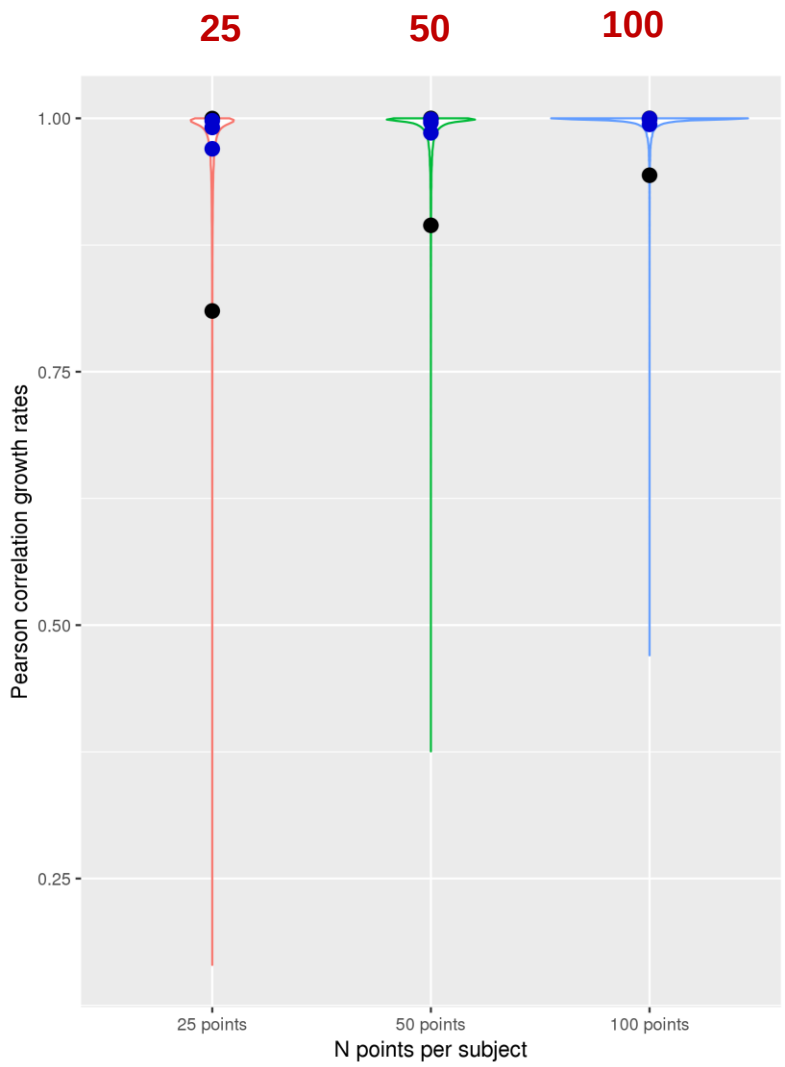
By number of subjects



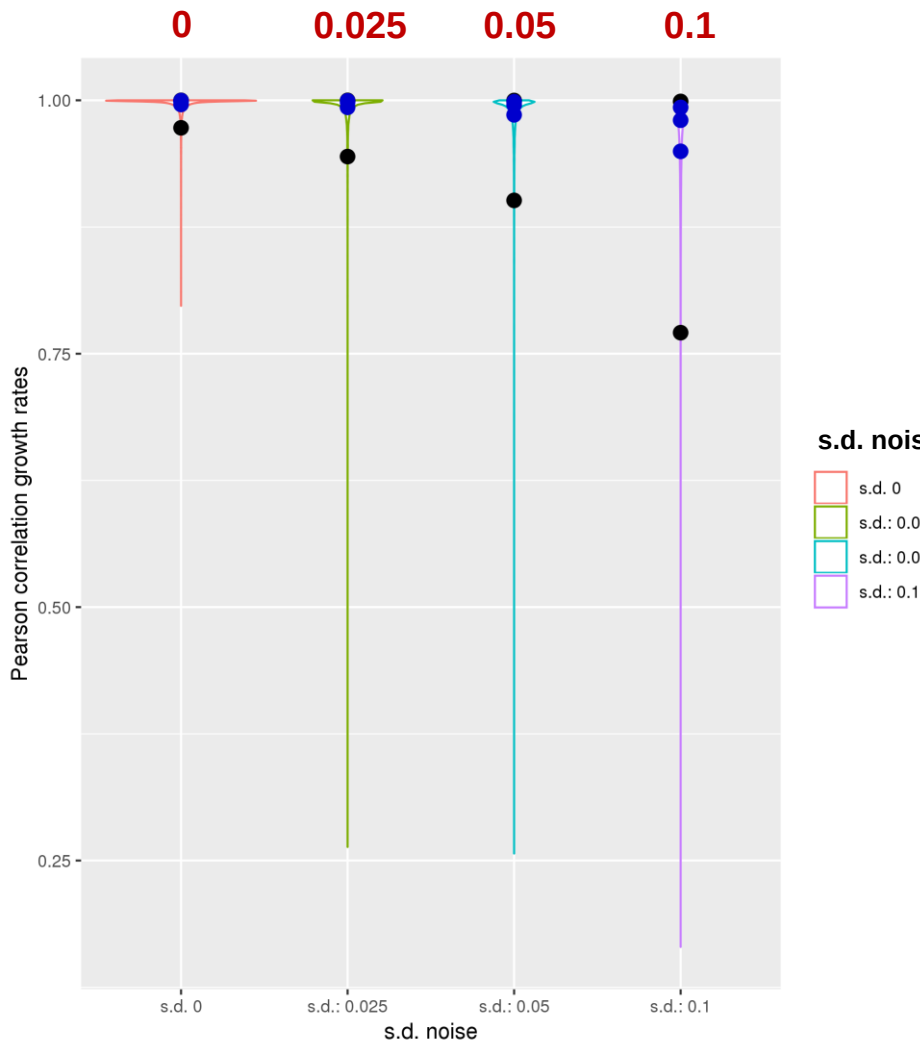
Blue dots: 25%, 50% and 75% percentiles
Black dots: 2.5% and 97.5% percentiles

Pearson correlation between estimated and original growth rates

By number of points per subject

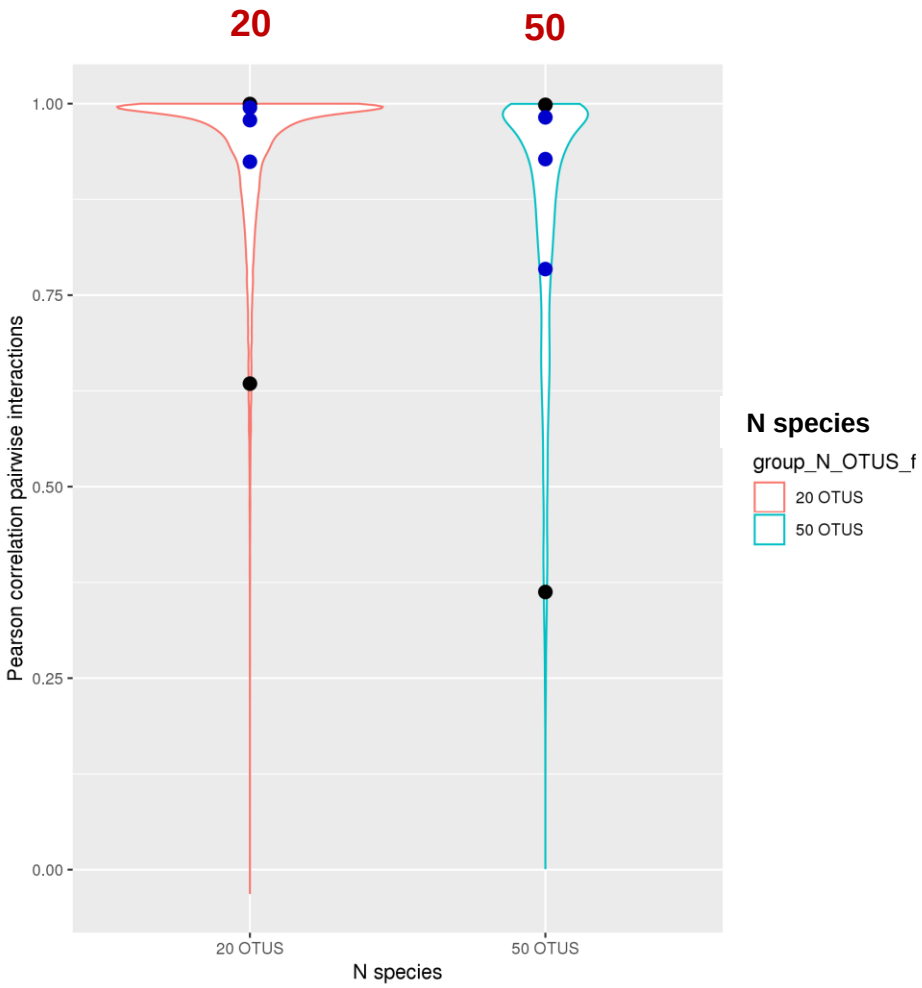


By standard deviation of noise

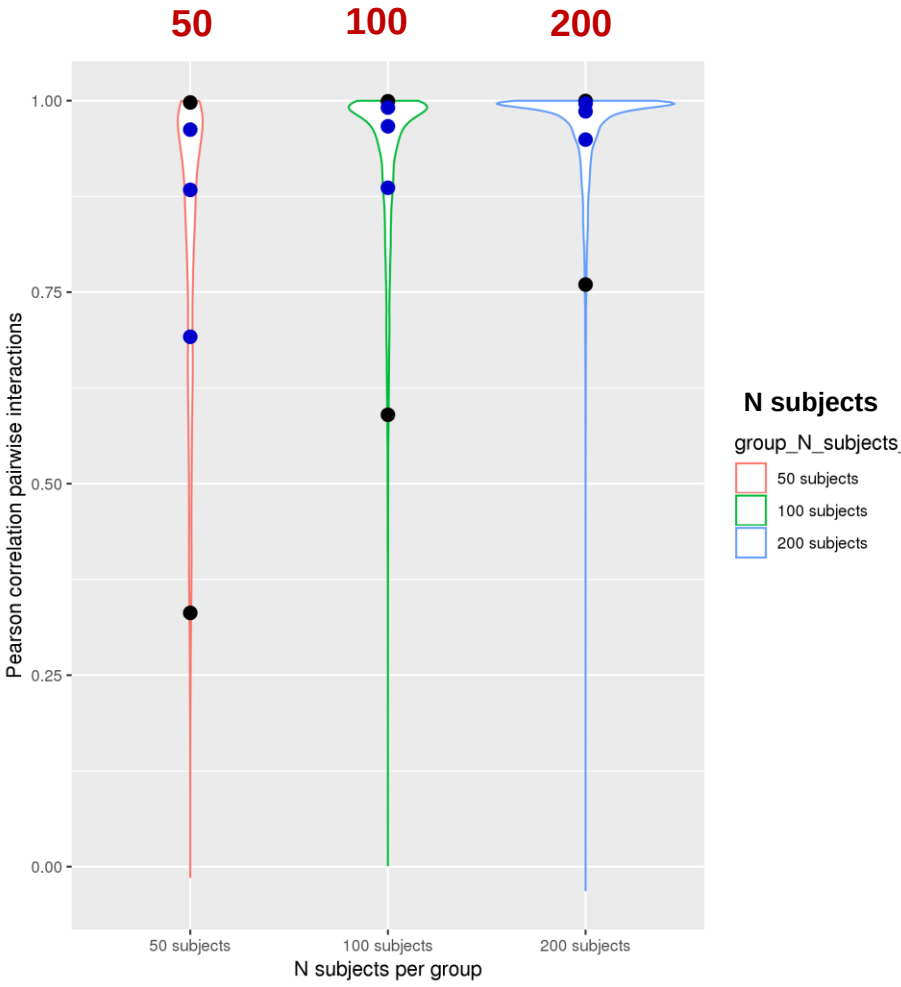


Pearson Correlation between estimated and original pairwise interactions

By number of species



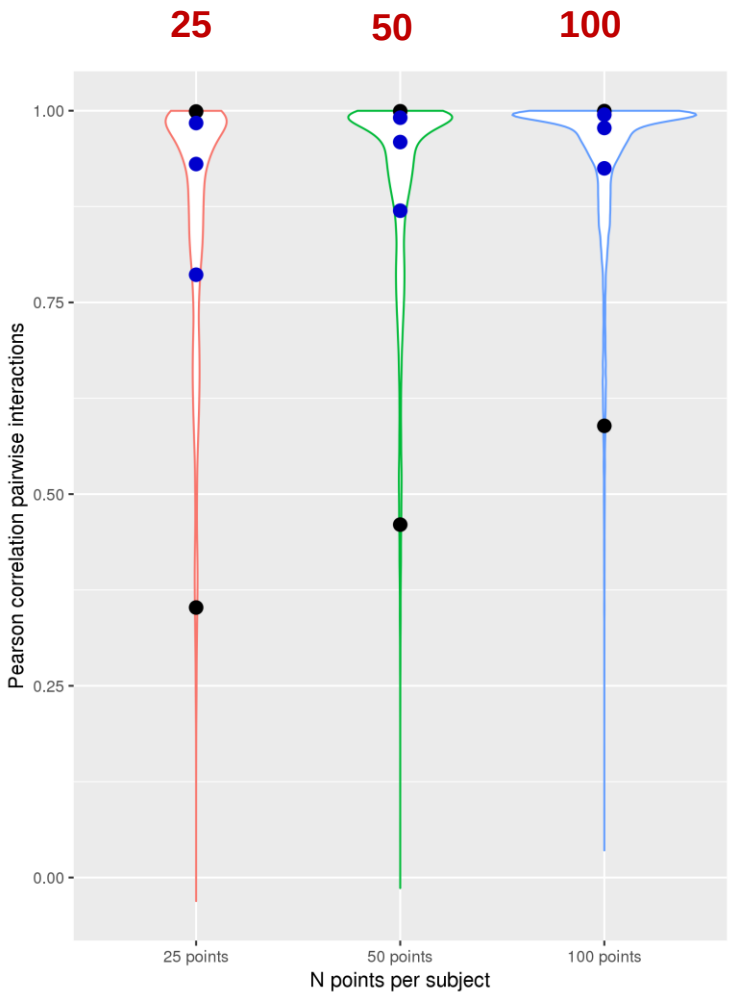
By number of subjects



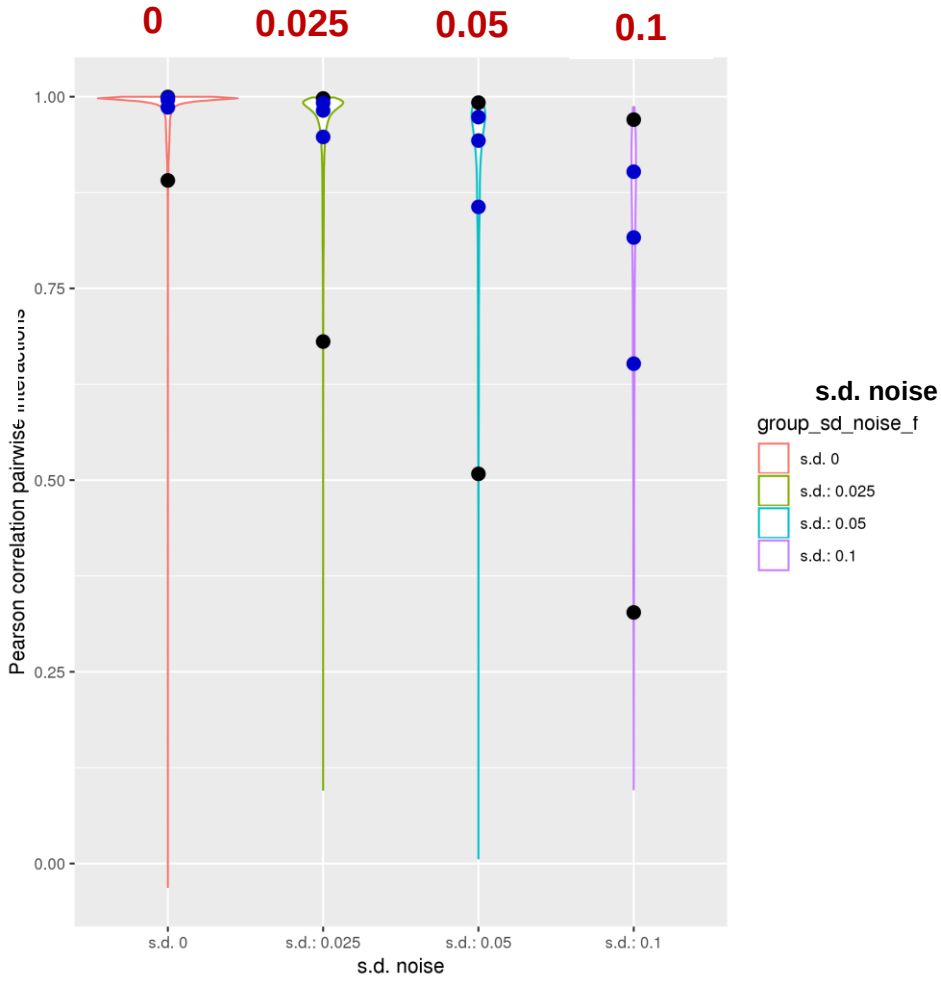
Blue dots: 25%, 50% and 75% percentiles
Black dots: 2.5% and 97.5% percentiles

Percentage of pairwise interactions with same sign

By number of points per subject



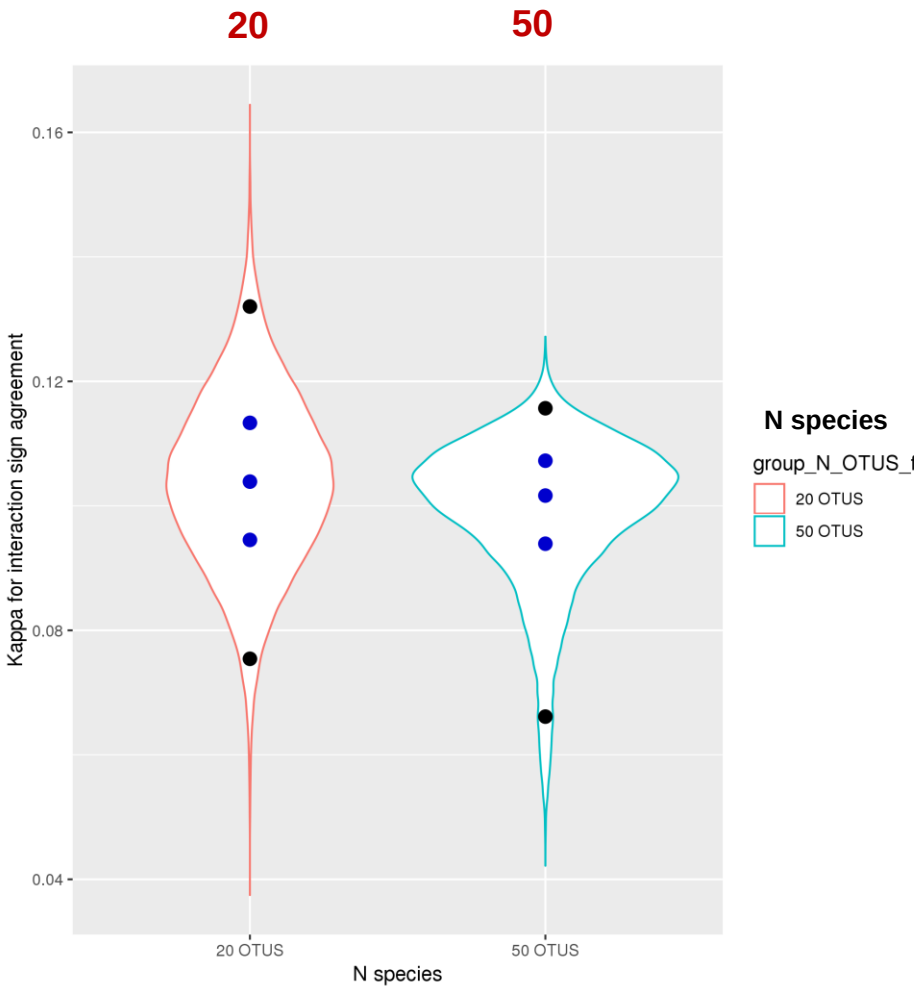
By standard deviation of noise



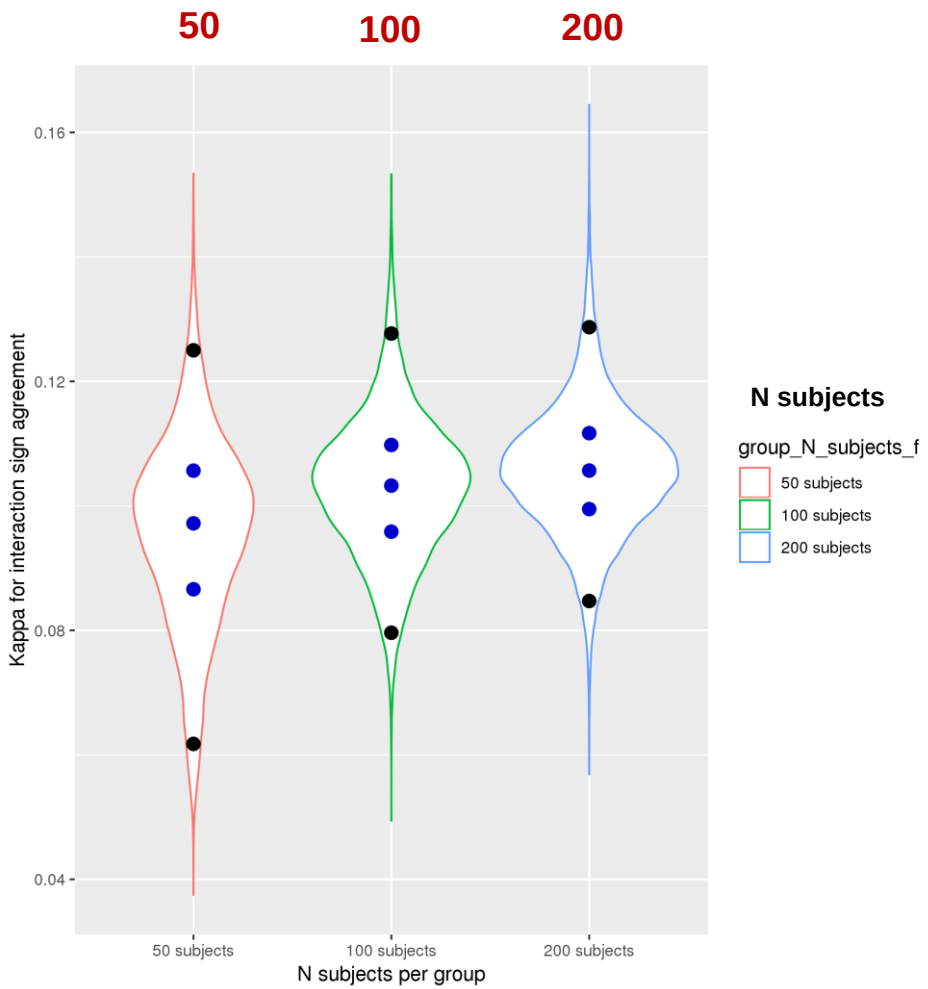
Blue dots: 25%, 50% and 75% percentiles
Black dots: 2.5% and 97.5% percentiles

Agreement in sign of interactions
Cohen's Kappa statistic

By number of species



By number of subjects

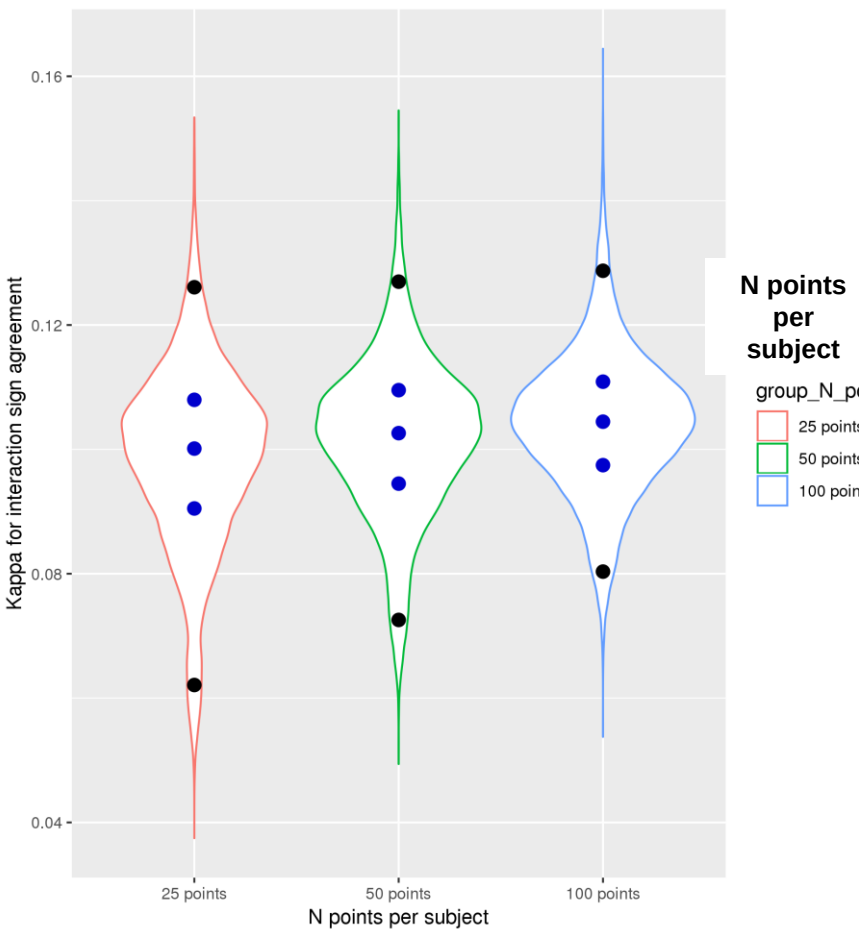


Blue dots: 25%, 50% and 75% percentiles
Black dots: 2.5% and 97.5% percentiles

Agreement in sign of interactions
Cohen's Kappa statistic

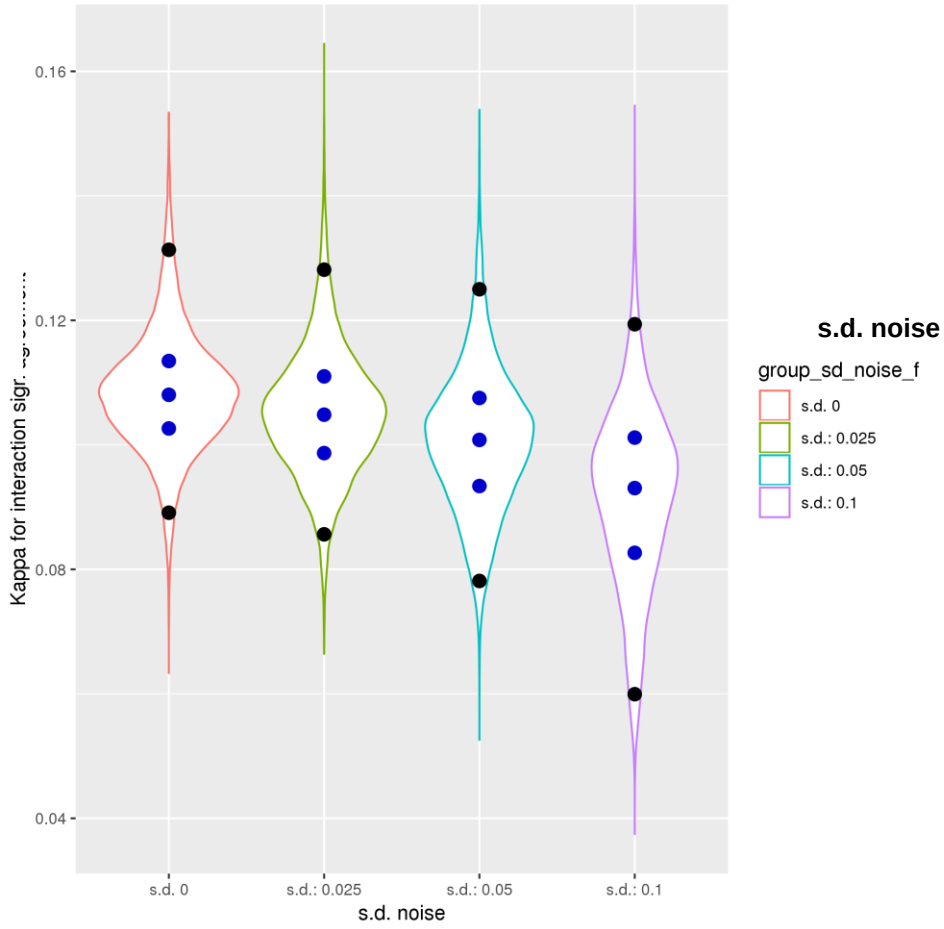
By number of points per subject

25 **50** **100**



By standard deviation of noise

0 **0.025** **0.05** **0.1**



Blue dots: 25%, 50% and 75% percentiles
Black dots: 2.5% and 97.5% percentiles

Smooth and integrate with cross-validation and "bagging"

- The estimates we obtain withe the S & I method are all non-zero, while we actually expect the interaction matrix to be pretty sparse.

- We therefore used the same method **with**
  - **cross-validation**, adding progressively more interaction parameters in the model, as long as it decreases the error (integral of sum of squares) enough, and
  - **"bagging":** several replicates, with final estimates taken to be the medians of the estimated parameters across the replications.

- This method has already been used with forward stepwise regression with the discretized GLV model.
  ("LIMITS" algorithm, Fisher Mehta, PLoS One, 2014).

Smooth and Integrate (S & I) with cross-validation and bagging

Method applied species by species.

- For each species and each "replicate":
  1. Divide the data set in a training set (for estimation) and a testing set.
  2. Start with a model with growth rate and self-interaction only
  3. Add 1 interaction parameter at the time in the model.
  4. With each added parameter, use the S & I method to estimate the parameters on training data set and compute integral of sum of squares with testing dataset.
  5. Add interaction parameter that gives lowest integrated error to model.
  6. If relative decrease in integrated error greater than specified threshold, return to Step 3 and add interaction parameters (1 at the time) among those not yet in model.

     Otherwise stop → final set of parameter estimates

Do this for each replicate.

Final parameters = medians across the replicates.

Comparison of S & I with and without cross-validation & bagging

Estimation with

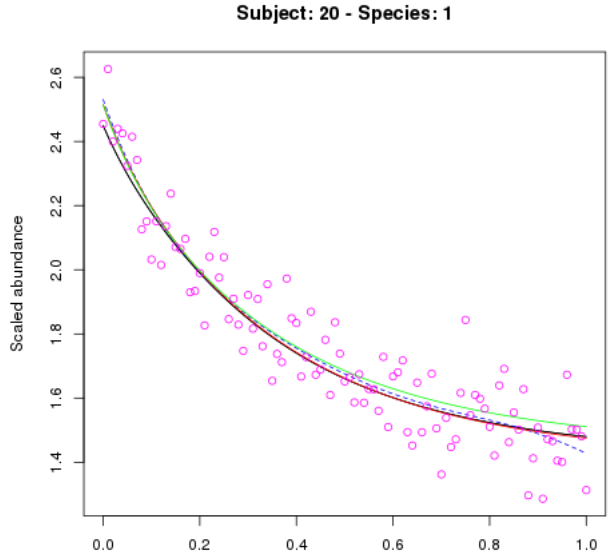
- "Smooth and integrate" (*S & I*)
- "Smooth and integrate" with cross-validation and bagging (*S & I + B*)

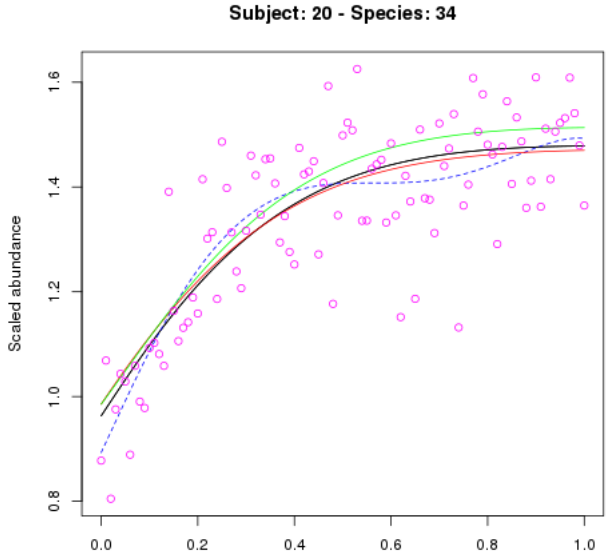
with

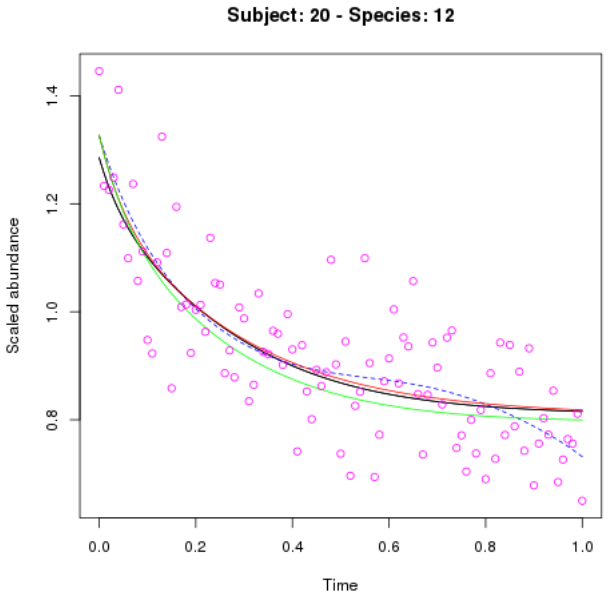
- 50 species
- 50 or 100 subjects
- 50 or 100 data points per subject
- Random noise Normal with a s.d. of 0.1

- 100 replications for bagging
- 2/3 of subjects for training set and 1/3 for testing set
- Stop adding interaction parameters when relative improvement < 2%

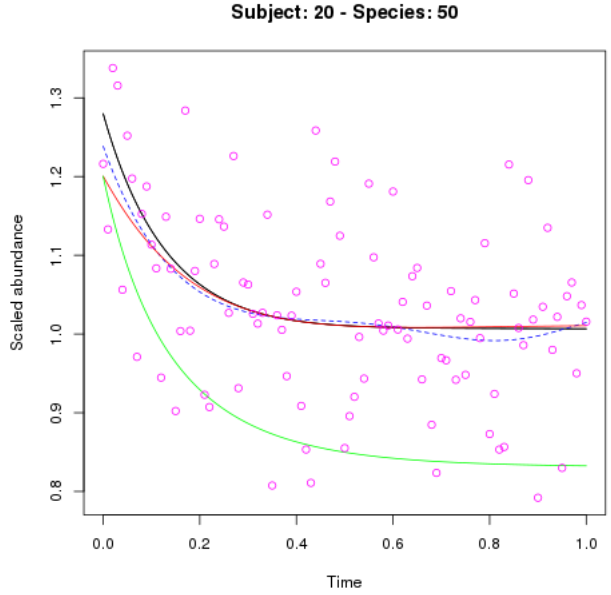
GLV model based on original vs. Estimated parameters
4 different species for a specific subject





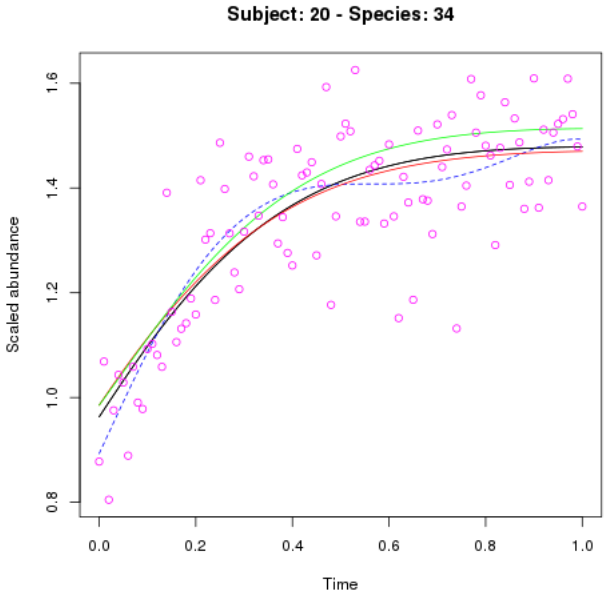


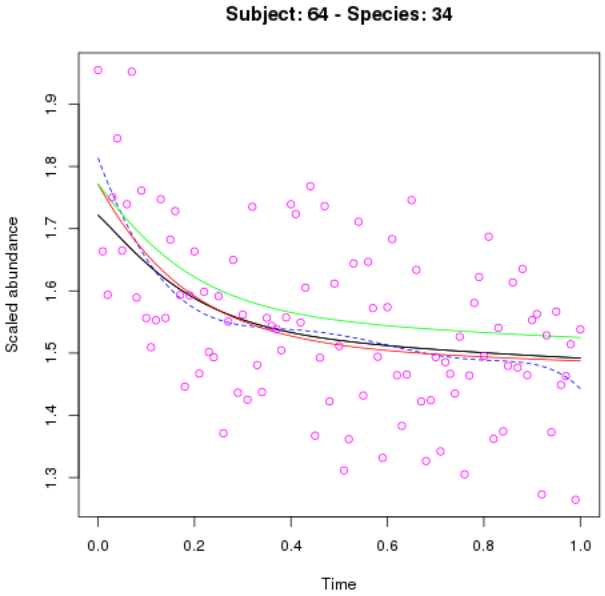
Model with 100 subjects and 100 points per subject

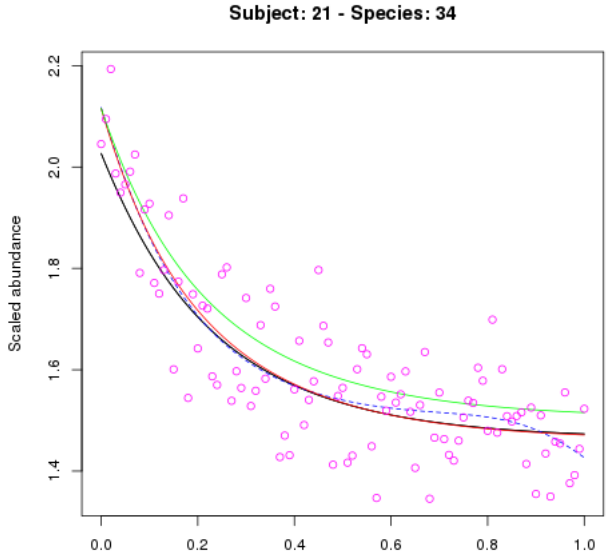


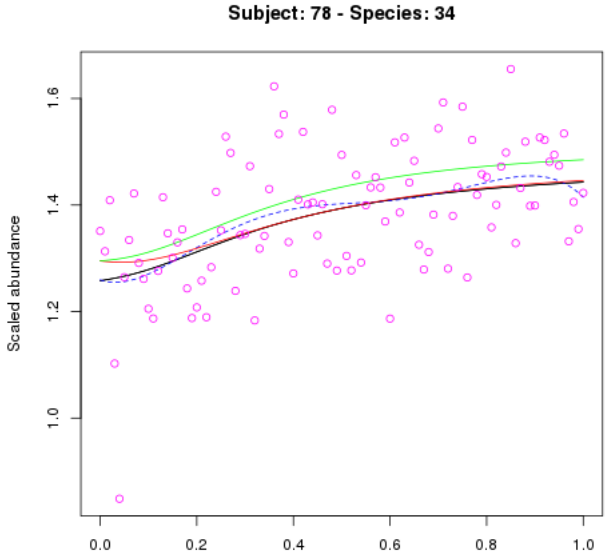
Black:original GLV with original parameters - Magenta: generated noisy data - Blue (dashed): local polynomials
Red: GLV with parameters estimated with S & I - Green: GLV with parameters estimated with S & I + B

GLV model based on original vs. Estimated parameters
4 different subjects for a specific species





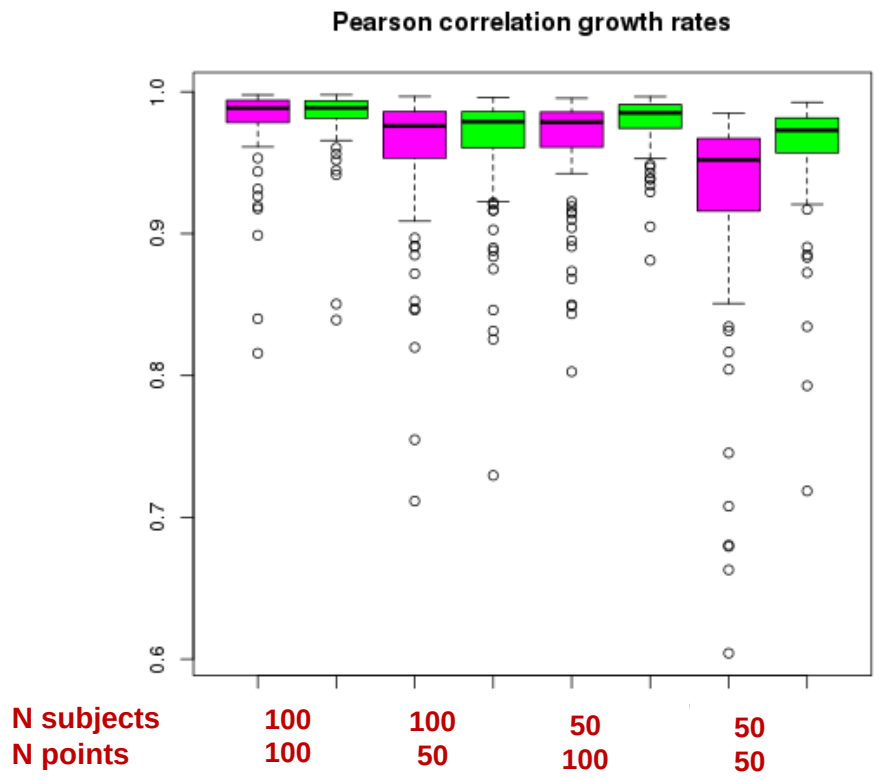




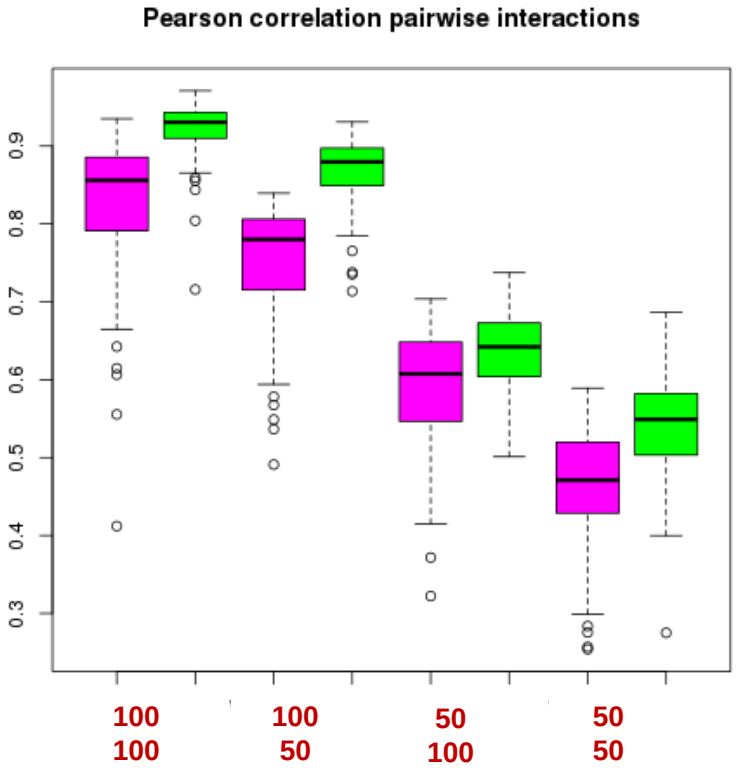
Black:original GLV with original parameters - Magenta: generated noisy data - Blue (dashed): local polynomials
Red: GLV with parameters estimated with S & I - Green: GLV with parameters estimated with S & I + B

Outcomes with S & I versus S & I + B

Pearson correlation
GROWTH RATES
Estimated vs. Original parameters



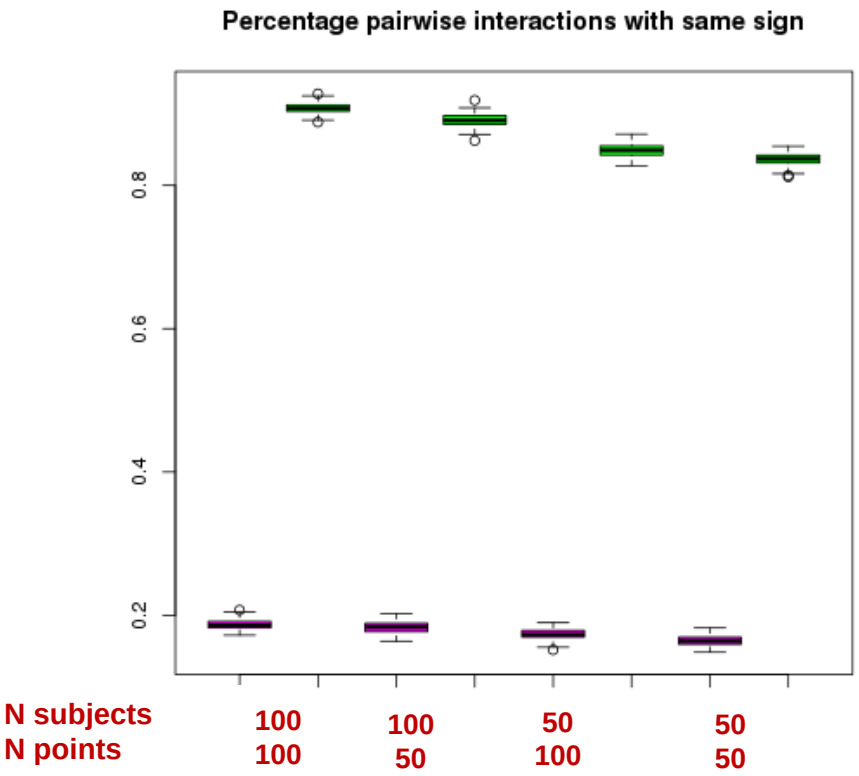
Pearson correlation
PAIRWISE INTERACTIONS
Estimated vs. Original parameters



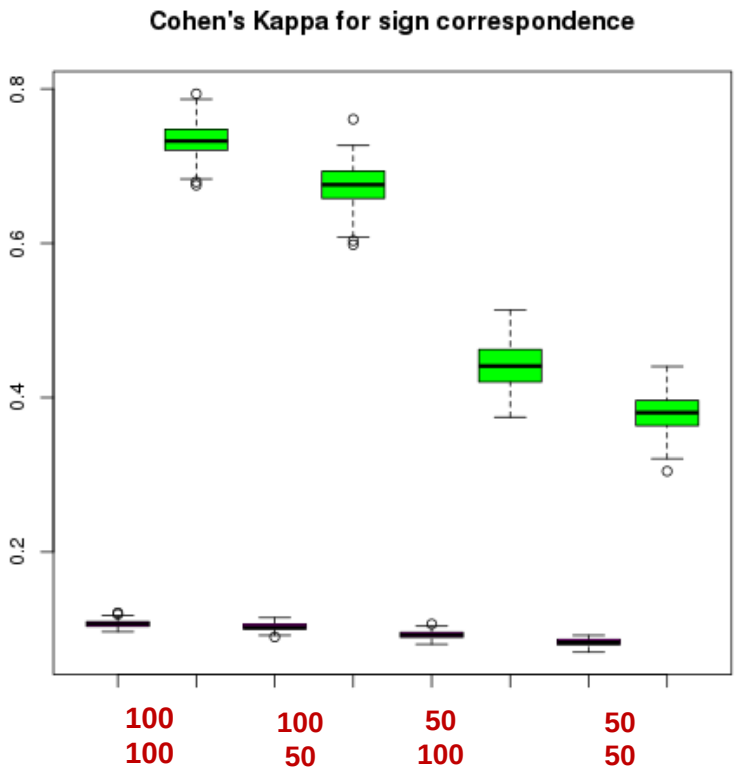
Method: **Magenta: S & I**
Green: S & I + B

Outcomes with S&I
with vs. without cross validation and bagging

PERCENT PAIRWISE INTERACTIONS WITH SAME SIGN
Estimated vs. Original parameters



COHEN's KAPPA
for aggreement of SIGN
Estimated vs. Original parameters



Method: **Magenta: S & I**
Green: S & I + B

Conclusions and challenges

- Understanding and quantifying the interactions between bacterial species important as they might affect interventions targeting the microbiome.
- Quantify interactions from longitudinal data is challenging.

- We propose a method to estimate the growth rates and the pairwise interactions using smoothing and direct integration, with cross-validation and bagging.
- Use of the method based in simulated data
  - helps to quantify the impact of different design characteristics on the estimated parameters
  - Indicates that valuable information about growth and interactions can be estimated with samples that are large enough even with rather substantial noise.

- There is a need to
  - generate more longitudinal data for microbiome
  - better understand and quantify the type and magnitude of variability of bacterial abundances over time.
- Simulations useful to inform design of studies about microbiome.