Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A sparse Bayesian hierarchical vector autoregressive model for microbial dynamics in a wastewater treatment plant[☆]

Naomi E. Hannaford[a], Sarah E. Heaps[c,*], Tom M.W. Nye[a], Thomas P. Curtis[b], Ben Allen[a], Andrew Golightly[c], Darren J. Wilkinson[c]

[a] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, United Kingdom
[b] School of Engineering, Newcastle University, Newcastle upon Tyne, United Kingdom
[c] Department of Mathematical Sciences, Durham University, Durham, United Kingdom

ARTICLE INFO

Article history:
Received 29 June 2021
Received in revised form 16 October 2022
Accepted 17 October 2022
Available online 17 November 2022

Keywords:
Regularised horseshoe
Shrinkage
Time series clustering
Microbial dynamics

ABSTRACT

Proper function of a wastewater treatment plant (WWTP) relies on maintaining a delicate balance between a multitude of competing microorganisms. Gaining a detailed understanding of the complex network of interactions therein is essential to maximising not only current operational efficiencies, but also for the effective design of new treatment technologies. Metagenomics offers an insight into these dynamic systems through the analysis of the microbial DNA sequences present. Unique taxa are deduced through sequence clustering to form operational taxonomic units (OTUs), with per-taxa abundance estimates obtained from corresponding sequence counts. The data in this study comprise weekly OTU counts from an activated sludge (AS) tank of a WWTP along with corresponding measurements of chemical and environmental (CE) covariates. Directly fitting a model to the OTU data is incredibly challenging because of the high dimensionality and sparsity of the observations. The first step is therefore to aggregate the OTUs into twelve microbial communities or “bins” using a seasonal phase-based clustering approach. The mean abundances in the twelve bins are assumed to vary over time according to a multivariate linear regression on the CE covariates. Deviations from the mean are then modelled using a vector autoregressive (VAR) model of order one, which is a linear approximation to the commonly used generalised Lotka-Volterra (gLV) model. Sparsity is assumed in the interactions between microbial communities by carrying out inference in a hierarchical Bayesian framework which uses a shrinkage prior for the autoregressive coefficient matrix of the VAR model. Different shrinkage priors are explored by analysing simulated data sets before selecting the regularised horseshoe prior for the biological application. It is found that ammonia and chemical oxygen demand have a positive relationship with several bins and pH has a positive relationship with one bin. These results are supported by findings in the biological literature. Several negative interactions are also identified. These novel biological findings suggest OTUs in different bins may be competing for resources and that these relationships are complex. Although simpler than a gLV model, the VAR model is still able to offer valuable insight into the microbial dynamics of the WWTP.

[☆] Additional text, plots, code and data related to this article is available as a supplementary material of the online version of the paper.

[*] Corresponding author.
E-mail address: sarah.e.heaps@durham.ac.uk (S.E. Heaps).

https://doi.org/10.1016/j.csda.2022.107659

1. Introduction

Due to recent advances in sequencing technology, there has been an increasing interest in longitudinal studies of microbial communities from a large range of environments. Unique ecological insights into response to perturbations (or environmental changes) and community stability can be gained from such studies (Faust et al., 2015). Furthermore, the complex non-linear interactions between different microbes result in many possible pseudo-stable states (Goyal et al., 2018). These interactions and others between different microbes and their environment contribute significantly to microbial dynamics (Konopka et al., 2015).

In microbial biology, DNA sequences extracted from environmental samples are grouped together into operational taxonomic units (OTUs) using a clustering algorithm, where typically a single OTU contains sequences that are at least 97% similar with each other (Bunge et al., 2014; Xia et al., 2018). Crucially, a particular OTU does not necessarily exactly correspond to a true biological species, but OTUs can be thought of as pragmatic proxies for classifying taxa. We have weekly OTU counts from the activated sludge (AS) of a UK-based wastewater treatment plant (WWTP) over five years, along with corresponding measurements of chemical and environmental (CE) covariates. After wastewater enters the WWTP, it undergoes the physical process of primary sedimentation, during which large solids are settled out. The wastewater that emerges from the primary sedimentation tank is called settled sewage and is fed into an aerated tank. The content of the aerated tank is the AS, which plays a pivotal role in wastewater treatment. Primarily responsible for the consumption of dissolved organic material, it comprises a plethora of different aerobic and anaerobic microorganisms (Shchegolkova et al., 2016).

Microbial communities within AS are complicated biosystems with a network of interconnected trophic links. For example, to degrade complex polymers, such as proteins, carbohydrates and lipids, many enzymes are required in a multi-stage process. Several species of microorganisms are needed for complete biodegradation. Gaining theoretical understanding of how these large biological systems work is likely to accelerate the creation of better biotechnological procedures (Curtis et al., 2003).

In the literature, community dynamics are often described by the generalised Lotka-Volterra (gLV) (Lotka, 1926; Volterra, 1926) differential equations, where changes in microbial counts are modelled as a function of taxon-specific growth rates and pairwise interactions. For example, Mounier et al. (2008) used a gLV model to identify interactions within a cheese microbial community. The gLV model is used to characterise the dynamics of a K-species system, where $K > 2$. Changes in population of species i are described by

$$\frac{d}{dt} y_i(t) = b_i y_i(t) + y_i(t) \sum_{j=1}^{K} a_{ij} y_j(t), \tag{1}$$

where $y_i(t)$ is the population size of species i at time t, b_i is the growth rate of species i in the absence of any competition and $\mathrm{A} = \left(a_{ij}\right)$ is a matrix of pairwise interactions.

Often in microbiome studies, the problem of finding the gLV parameters is simplified by using (1) to express $\frac{d}{dt} \log y_i(t)$ as a linear function in the $y_j(t)$, and then discretising the left hand side to view the problem of estimating the gLV parameters as one of least squares (Stein et al., 2013; Fisher and Mehta, 2014; Bucci et al., 2016). Despite widespread use of gLV models, Gibbons et al. (2017) investigated microbial dynamics in the human gut with a sparse vector autoregressive (VAR) model. This model offers the advantage over gLV models of allowing quantification of uncertainty by explicitly modelling error. A key difference between the two approaches is that VAR models assume linear dynamics, whereas gLV models assume non-linear dynamics.

1.1. Scientific questions, contributions and organisation of the paper

Consider a VAR model of order one (VAR(1)) for a K-dimensional time series,

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \mathrm{A}\left(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{y}_t$ is a K-dimensional vector of observations at time t, A is a $K \times K$ matrix of autoregressive coefficients, $\boldsymbol{\mu}$ is a K-dimensional location parameter and $\boldsymbol{\epsilon}_t$ is a vector of K normally distributed errors at time t, that is $\boldsymbol{\epsilon}_t \sim \mathrm{N}_K(\mathbf{0}, \Sigma)$. As we show in Section S1 of the Supplementary Materials, a VAR(1) model can be regarded as a linear approximation to the non-linear numerical solution of a stochastic Lotka-Volterra system. Crucially, linearised versions of the interaction terms are preserved which, when non-zero, allow lagged cross-dependence between the time series and relate closely to the concept of Granger causality (Granger, 1969). Although the second-order non-linear interaction terms are omitted, linearisation of the stochastic Lotka-Volterra system affords substantial practical benefits by greatly simplifying the process of model-fitting. The focus of this paper is the development of a novel model for the WWTP microbial abundances that is based on a VAR(1) process. By modifying (2) to allow the location parameter to change over time, we are able to additionally incorporate information from the CE covariates. By fitting this model to the WWTP data we seek to answer two main questions: whether a VAR(1) model can provide useful biological insights into microbial interactions and whether there is evidence of a relationship between the CE covariates and microbial abundance.

Table 1
Chemical and environmental covariates.

Covariate	Unit
Ammonia, Chloride, COD, DO, Fluoride, MLSS, MLVSS, Nitrate, Nitrite, Phosphate, Sulphate	mg/L
pH	pH
Temperature	Celsius

As explained in Section 3, we address the problem of high-dimensionality in the OTU data by using a novel seasonal phase-based clustering approach to form 12 “bins” of OTUs. As a result, the bins have a circular time-ordering, which means that a particular bin k typically has its peak abundance one month before bin $k-1$ and one month after bin $k+1$. Therefore, it is unlikely that the previous abundances for all bins influence the current abundance of any particular bin once abundance in the neighbouring bins is known. A sparse autoregressive matrix containing many zeros reflects this idea. We investigate different shrinkage priors for this parameter and compare the performance of the priors and their corresponding inferential procedures via a simulation study, before selecting a regularised horseshoe prior. As a by-product of this approach, we describe a novel extension to the work of Piironen and Vehtari (2017) on prior specification in sparse linear models, thereby providing a principled methodology for problems in which the response variable is vector-valued, rather than univariate. This involves careful specification of the hyperprior for the global shrinkage parameter using prior beliefs about sparsity.

The remainder of the paper is structured as follows. In Section 2, we describe the data and present findings from an exploratory analysis. Section 3 discusses clustering methods and describes a seasonal phase-based approach. In Section 4, we give a full model specification, followed by an exploration of shrinkage priors for the matrix of autoregressive coefficients in Section 5. In Section 6, the details of Bayesian inference for our model are given. Section 7 presents the results of applying our model to the WWTP data. Finally, we summarise our findings in Section 8.

2. Exploratory analysis

In this section, we describe the data and report the findings of an exploratory data analysis, which helps guide the adjustments that we make to the VAR(1) model in (2). We consider the datasets separately before analysing the combined data.

2.1. Data description

We have weekly counts of 9044 different OTUs measured at $N = 257$ time points, starting from 1st June 2011. Each year a week is missing for the Christmas period, which we treat as missing at random; the manner by which this assumption is incorporated in our model is explained in Section 4.3. The dimensions of the data present a significant inferential challenge. Given that the number of OTUs is much larger than the number of time points, fitting joint models to the counts of all OTUs would be computationally prohibitive. Furthermore, it is unclear whether the number of time points would be sufficiently large to detect interactions and there is the common issue of sparsity in the OTU table, that is, the presence of many zeros. In our data, approximately 90% of the counts are zero. These could correspond to OTUs that enter the system randomly and die out quickly. However, it is more likely that the zeros can be attributed to insufficient sampling depth. These two issues are tackled using clustering in Section 3.

Accompanying the OTU table is a taxonomy table containing the kingdom, phylum, class, order, family and genus of every OTU. However, some of the taxonomic ranks for many of the OTUs are missing. For example, roughly 54% of OTUs do not have a genus assigned. Table S2 in the Supplementary Materials shows the proportions of missing data for each taxonomic rank. The Ribosomal Database Project (RDP) classifier (Wang et al., 2007) was used for classification of each OTU. An OTU was classified as `NA` if its sequence was not sufficiently similar to entries in the RDP database.

Finally, we have measurements for 13 different CE covariates, shown in Table 1. Chemical oxygen demand (COD) measures the amount of oxidisable organic matter dissolved in the sample. Dissolved oxygen (DO) is the concentration of dissolved oxygen. Mixed liquor suspended solids (MLSS) is the concentration of suspended solids in the tank, determined by filtration and drying at a relatively low temperature. The related mixed liquor volatile suspended solids (MLVSS) records the mass of volatile material lost (evaporated) by heating filtered solids at a higher temperature. Supplementary Table S3 shows the proportions of missing data for each covariate. We describe how we account for the small amount of missing covariate data in Section 4.3.

In the sections that follow, we discuss an exploratory data analysis of the OTU table and taxonomy table. We then look at all the data together, with a particular focus on finding possible relationships between some of the CE covariates and the OTUs.

2.2. OTU data

A time series plot (Supplementary Fig. S1) of the total number of OTUs recovered shows an absence of trend. The total number of OTUs recovered denotes the total number of OTUs (sequences) detected in a sample, at a particular time point.

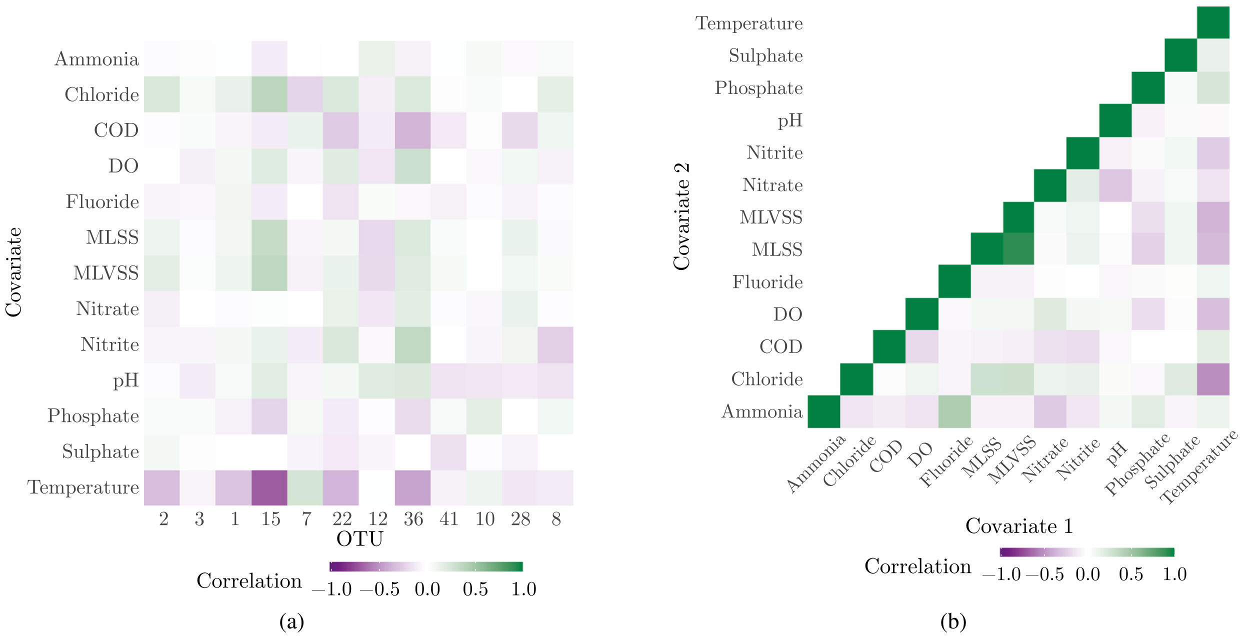

Fig. 1. Heatmaps of the pairwise correlations (a) between the top 12 OTUs and the CE covariates and (b) among the chemical and environmental covariates. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

Since there are 9044 different OTUs, it is not possible to analyse every single OTU, so instead we look at the 12 most abundant OTUs based on median abundance. Fig. S2 in the Supplementary Materials shows time series plots for these top 12 OTUs. OTUs 8, 15 and 28 clearly demonstrate seasonality with peaks occurring roughly once a year. OTU 1 also appears to have annual peaks, which are more visible when plotted on the log-scale (see Supplementary Fig. S3a). Seasonality with annual peaks is also evident from the stacked bar plot (shown in Supplementary Fig. S4).

2.3. Taxonomy table

Table S1 in the Supplementary Materials shows the unique numbers of kingdoms, phyla, classes, orders, families and genera. Supplementary Fig. S5 shows the time series plot for the top 12 genera based on median abundance. Missing genera, which are grouped together in a single “Unknown” group, represent a large proportion of the total abundance. *Rhodobacter* clearly shows seasonality with annual peaks in late February/early March. *Flavobacterium*, *Ferruginibacter* and *Trichococcus* also seem to display seasonality, although for the latter the seasonality is clearer on the log-scale (shown in Supplementary Fig. S3b). Supplementary Fig. S6 shows that the 12 genera account for about 60% of the total abundance on average at each time point.

Fig. S7 in the Supplementary Materials shows the time series plot for the 12 most abundant classes based on median abundance. There are hints of seasonality in some of the classes, even at this fairly coarse taxonomic rank, for example, *Flavobacteriia*, *Actinobacteria* and *Deltaproteobacteria* show rough annual peaks. *Clostridia*, *Bacilli* and *Gammaproteobacteria* also show seasonal behaviour. The most abundant class is *Alphaproteobacteria* and its time series profile is very noisy without any obvious annual peaks. *Alphaproteobacteria* form one of the most abundant groups of bacteria on the planet and are extremely diverse (Williams et al., 2007), so it is unsurprising that this class is the most abundant. The diversity of *Alphaproteobacteria* is also reflected here, as there are 1238 different OTUs from the class present. A class as diverse as this may have species that prefer different conditions and hence have peaks in population size at different times of the year. The stacked bar plot for the top 12 classes (Supplementary Fig. S8) shows the dominance of *Alphaproteobacteria* and *Betaproteobacteria*.

2.4. Analysis of the combined data

In this section we investigate relationships between OTU abundance and the CE covariates. We identify potential relationships between the CE covariates and the 12 most abundant OTUs, genera and classes. For each OTU, we can compute the relative abundance at each time point as the count for that OTU divided by total number of OTUs recovered at that time point. Fig. 1a shows a heatmap of the correlations between the covariates and the relative abundances of the top 12 OTUs. Some of the most abundant OTUs appear to be correlated with temperature. There appear to be some correlations between some of the OTUs and chloride, nitrite, COD, DO, phosphate, MLSS and MLVSS. Fig. 1b shows a heatmap of the pairwise correlations between the CE covariates. It seems that these covariates are potentially correlated with temperature. However,

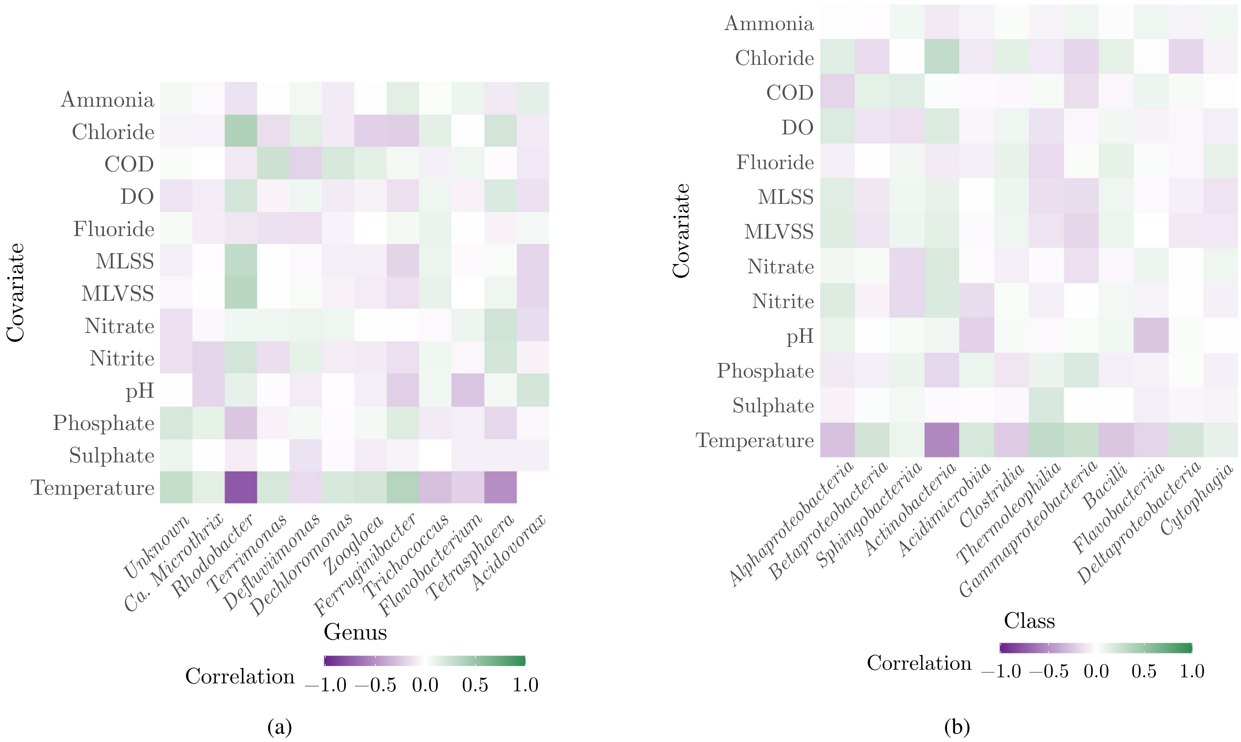

Fig. 2. Heatmap of the correlations between the chemical and environmental covariates and (a) the 12 most abundant genera and (b) the 12 most abundant classes. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

with the exception of temperature and chloride, these correlations are fairly weak, so it would be naïve to attribute all of the possible relationships that we see here to an indirect relationship with temperature. Microorganisms in AS are very diverse and interact with, feed on and utilise chemical compounds in many different ways. Nevertheless, it does seem that in general for these top 12 OTUs, if the correlation with temperature is weaker, then the correlations with other covariates tend to be weaker too.

Fig. 2a shows the heatmap of correlations between the covariates and the 12 most abundant genera. Most of the genera have at least a weak correlation with temperature, with the exception of *Acidovorax*, *Ca. Microthrix* and *Defluviimonas*. *Rhodobacter*, *Ferruginibacter* and *Tetrasphaera* appear to have a strong correlation with temperature, which we already identified as showing seasonal behaviour in Section 2.3. We also see that pH has a weak correlation with *Acidovorax*, *Ferruginibacter* and *Flavobacterium*. There are weak correlations present between: nitrite and *Tetrasphaera*; nitrite and *Rhodobacter*; and nitrate and *Tetrasphaera*. COD and chloride have weak correlations with a few of the genera.

A heatmap of the correlations between the top 12 classes and the CE covariates is shown in Fig. 2b. Most classes exhibit at least weak correlation with temperature. *Flavobacteriia* appears to have a negative correlation with pH and *Acidimicrobiia* appears to have a weak negative correlation with pH too.

2.5. Summary

From this exploratory analysis, we have identified signs of seasonality and the absence of time trend. Seasonality is taken into consideration via our time series clustering method in Section 3.1 and through the use of harmonic regression to model a time varying mean in the VAR(1) model in Section 4.2. Furthermore, we have identified relationships between some of the CE covariates and the relative abundances of the top 12 OTUs. These relationships persist even at the coarser taxonomic ranks of genera and classes. Thus, CE covariates are incorporated through linear regression in the time varying mean in Section 4.2.

3. Clustering

The data are counts of OTUs, where some OTUs have counts in the thousands and others have (mostly) counts of zero throughout time. Zeros can arise for structural reasons (“hard zeros”) or due to lack of sampling depth (“soft zeros”) (Kaul et al., 2017). As such, we would expect an excess of zeros over Poisson variation. Indeed, 91.5% of the counts are zero. Therefore, a natural approach might be to use a time series model for zero-inflated multivariate count data, for example, see Lee et al. (2018). However, there are over 9000 OTUs, indicating our model would have to allow over 81 million pairwise interactions. To make model-fitting more manageable we instead choose to cluster the data, following analyses by other authors (see Eiler et al. (2012); Stein et al. (2013); David et al. (2014); Dam et al. (2016)). Choosing a small enough number of clusters removes the complication of zero inflation and allows us to make the simplifying assumption that our data can be modelled as continuous.

An approach that reduces the dimensionality of the data, but also retains all of the OTUs, is taxonomy-based clustering. This involves taking the n most abundant taxa at each time point that represent a high percentage, say 90%, of the total abundance and grouping the remaining taxa into an "other" category. The term taxa here could refer to any taxonomic rank. For example, Stein et al. (2013) grouped OTUs into the top ten genera and an "other" category in their work to infer gut microbiota ecology in mice. This approach is unsuitable for our data because of the large proportion (54.1%) of missing taxonomic information at the genus level and the large number of genera (187) required to capture 90% of the abundance. Even when considering median abundance and coarser taxonomic ranks, we find that the finest taxonomic rank we can use without having an unknown as a group is class, which is possibly too coarse. As we noted in Section 2.3, *Alphaproteobacteria* was the most abundant class in the AS tank (with 1238 different OTUs) but this class is known to be extremely diverse in general (Williams et al., 2007). Modelling the change in its abundance over time and its interactions with other classes and the environment is unlikely to yield biologically useful insight, given that the different OTUs within the class may prefer different conditions.

Dam et al. (2016) researched dynamic models of the complex microbial metapopulation in a lake and suggested that, for characterising interaction dynamics, clustering by taxonomy is not an effective strategy. They proposed an alternative method of clustering OTUs, where they define peak profiles, which involves identifying positions in time where each OTU has its largest abundance(s). The OTUs are then clustered into "subcommunities" based on these profiles with remaining OTUs placed in an additional group. The rationale is that these subcommunities represent OTUs with similar dynamics perhaps because of symbiotic relationships or shared dependence on the environment. Since we have clear evidence of seasonality in the WWTP data, we adopt a similar approach.

3.1. Time series clustering

We implement our time series clustering approach as follows. OTU counts are available for 51 weeks every year (Christmas week is excluded) for each of five consecutive years. For every OTU, we calculate the 51 weekly means, in each case averaging the observations for that calendar week across years 1 through 5. The 51 weekly means are then scaled to have standard deviation equal to one. Next, the scaled weekly means for each OTU are transformed from a set of 51 coefficients on a Euclidean basis to a set of 51 coefficients on a Fourier basis, with 25 harmonics of frequency $k = 1, \dots, 25$ cycles per year, and an intercept. The first harmonic dominated in amplitude in most cases which verified that many of the OTUs follow a simple annual cycle. This allowed a more automatic version of the peak profiles approach of Dam et al. (2016) to be implemented by clustering based on the phase of the first harmonic. Denote the phase of the first harmonic for OTU i by ϕ_i. The interval $[-\pi, \pi]$ is divided into 12 equally sized intervals and we assign OTU i to the interval in which ϕ_i lies for all i. This gives 12 clusters, which we call "bins". Let $\tilde{w}_{ti}$ be the count of OTU i at time point t. The set of OTUs in bin j is S_j and $w_{tj} = \sum_{i \in S_j} \tilde{w}_{ti}$ is the count for bin j at time t. Visual inspection of the counts of each bin reveals that the bins peak once per year, with different bins peaking in different months.

To stabilise the variance of the w_{tj} series over the year we log-transform the counts of the bins and set $\tilde{y}_{tj} = \log(w_{tj})$. We then scale the log counts of each bin so that their variance is roughly one (see Section 5.2), denoting by $y_{tj} = \tilde{y}_{tj}/\bar{s}$ for all t and $j = 1, \dots, 12$, where $\bar{s} = (\sum_{j=1}^{12} s_j)/12$, s_j is the standard deviation of $\tilde{y}_{1:N,j}$, y_{tj} denotes the scaled log counts and N is the number of time points. Fig. 3 shows the time series plots of the scaled log counts for the 12 bins. The plots do not seem to demonstrate any signs of a time trend, which could indicate a mean net growth rate of zero. Each bin clearly shows seasonal behaviour with a peak every year, with the exception of bins 2 and 3, where the peaks are not as obvious. We can see that for each bin the annual peaks are different. Bin 1 seems to peak in February, bin 2 seems to peak in January, bin 3 seems to peak in December and so on. This labelling of the bins is a consequence of the first time point in the series being at the start of the June and the interpretation of the phase for each bin. For $\phi = 0$ we have a harmonic (sine wave) that is 0 at time $t = 0$, which roughly corresponds to the last week of May. The peak of this harmonic will be at the end of August. Now, for example, take bin 1, which contains all OTUs with $\phi \in [-\pi, -5\pi/6)$. This corresponds to the harmonic being shifted five to six months forward in time (to the right) and means that OTUs in this bin typically peak anywhere between the end of January and February.

4. Model description

As discussed in Section 1.1, a vector autoregression is chosen as the model for the clustered OTU data. The novelty in our approach lies in the use of a hierarchical prior for the autoregressive coefficient matrix and error variance matrix. This is constructed to allow sparsity in the former whilst facilitating the development of a principled methodology for the specification of the hyperparameters in its prior. Before discussing the prior for the autoregressive coefficient matrix (Section 5), we describe our model in this section.

Our model has the high level structure $\boldsymbol{y}_t = \boldsymbol{\mu}_t\{\mathrm{CE}_t\} + \mathrm{A}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_{t-1}\{\mathrm{CE}_t\}) + \boldsymbol{\epsilon}_t$, where $\{\mathrm{CE}_t\}$ represents the incorporation of the CE data into the time varying mean $\boldsymbol{\mu}_t$ and $\boldsymbol{\epsilon}_t \sim \mathrm{N}(0, \Sigma)$ are error terms. First, we describe our choice for the

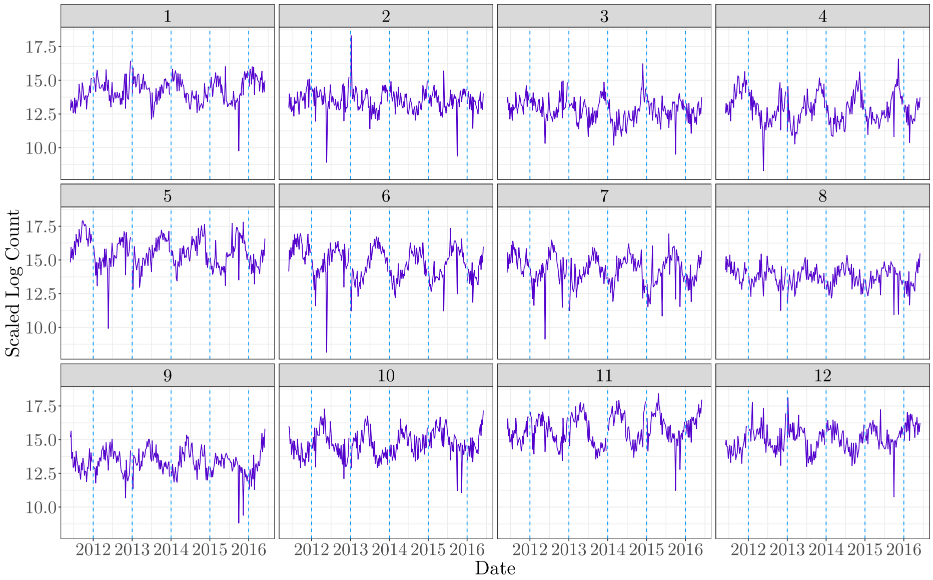

Fig. 3. Time series plots of the scaled log counts for the 12 bins. The dashed vertical lines represent the start of each year.

parametric structure of the variance matrix Σ for the error terms $\boldsymbol{\epsilon}_t$. Then we discuss how the time varying mean is constructed to allow for seasonality and the effects of the CE covariates. Finally, we conclude this section by describing how missing data are handled.

4.1. Error structure

When modelling with a vector autoregression, it is common to adopt a parsimonious parametric form for the error variance matrix Σ, for example, by assuming that Σ is diagonal, that is $\Sigma = \sigma^2 \mathrm{I}_K$, where I_K represents a $K \times K$ identity matrix. Since the clustered OTU data have a circular time-ordering, with the OTUs in neighbouring bins (or bins 12 and 1) typically peaking in abundance in neighbouring months, a more appropriate parametric form is to assume Σ^{-1} is a symmetric, circulant, tridiagonal matrix of the form

$$\Sigma^{-1} = \begin{pmatrix} \sigma_0^{-2} & \omega & 0 & 0 & \cdots & 0 & 0 & 0 & \omega \\ \omega & \sigma_0^{-2} & \omega & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \omega & \sigma_0^{-2} & \omega & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \omega & \sigma_0^{-2} & \omega \\ \omega & 0 & 0 & 0 & \cdots & 0 & 0 & \omega & \sigma_0^{-2} \end{pmatrix}. \tag{3}$$

This is the precision structure of a circular first-order autoregressive model (Gelfand et al., 2010, Chapter 13) for which the interpretation of the errors, at any time $t \in \{1, \ldots, N\}$ is as follows: (i) for $j < k$, the errors of y_{tk} have stronger correlations with the errors of y_{tj} when $\min\{k - j, j + K - k\}$ is smaller; (ii) the correlation between the errors of y_{tj} and $y_{t,j+\ell}$ is the same as the correlation between the errors of y_{tj} and $y_{t,j-\ell}$, where $j + \ell$ and $j - \ell$ are in arithmetic modulo K (and 0 is written as K rather than 0).

To ensure that Σ^{-1} is positive definite, it is convenient to reparameterise in terms of $\sigma_0^{-2} = (\varpi_0 + \varpi_1)/\sqrt{2}$ and $\omega = (\varpi_0 - \varpi_1)/2\sqrt{2}$. The precision matrix is then positive definite if and only if $\varpi_i > 0$ for $i = 0, 1$. The full derivation of this is given in Section S3 of the Supplementary Materials.

4.2. Time varying mean

Due to the manner in which the OTU data were clustered, it is likely that a time varying mean will be more appropriate than a mean which is static over time. This was also evident from the time series plots of the bins in Fig. 3. We therefore modify the simple VAR(1) model in (2) to give

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \mathrm{A}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_{t-1}) + \boldsymbol{\epsilon}_t, \tag{4}$$

for $t = 2, \ldots, N$, where A is unstructured and $\boldsymbol{\epsilon}_t$ is normally distributed with zero-mean and a precision matrix of the form (3). To capture the seasonal variation of each bin, we use a harmonic regression to fit a time varying mean, that is

$$\boldsymbol{\mu}_t = \boldsymbol{\alpha}_t + \sum_{j=1}^{J} \left\{ \boldsymbol{\beta}_j \sin\left(\frac{2\pi tj}{52}\right) + \boldsymbol{\gamma}_j \cos\left(\frac{2\pi tj}{52}\right) \right\}, \tag{5}$$

where J is the number of harmonics. After fitting the model with $J = 1, \ldots, 4$ harmonics, we select $J = 2$ for our final model because there was little evidence of $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ being non-zero for $j = 3, 4$. This assessment was based on the marginal posteriors for the β_{jk} and γ_{jk} for $j = 3, 4$ and $k = 1, \ldots, K$, each of which assigned near equal probability to either sign. Thus, in our final model, $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ correspond to the harmonic which repeats once per year and $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}_2$ to the harmonic which repeats twice per year. We then capture the information from the CE covariates by expressing the intercept term $\boldsymbol{\alpha}_t$ at time t as a linear combination of the CE covariates at the previous time point $t-1$. This was based on expert biological judgement, motivated by the idea that the effect of any environmental conditions is unlikely to be instantaneous.

Let $\tilde{\mathrm{X}}$ be a $N \times L$ matrix of our CE covariates. We find the covariate data are skewed and so apply a square-root transformation. This transformation relates to how we handle missing data (see Section 4.3). Each column (covariate) is then standardised, denoting the resulting matrix as X. We let

$$\alpha_{tk} = b_{0k} + b_{1k}x_{t-1,1} + \cdots + b_{Lk}x_{t-1,L}, \tag{6}$$

where $x_{t\ell}$ is the measurement of covariate ℓ at time t and $b_{\ell k}$ is a regression coefficient for bin k and covariate ℓ, noting that b_{0k} is the intercept term for bin k. We collect the b_{ij} into a $(L+1) \times K$ matrix B in which b_{ij} is the $(i+1, j)$-element.

To select which covariates to include, we fit the model without any CE covariates so that $\alpha_{tk} = b_{0k}$ for $t = 2, \ldots, N$ and $k = 1, \ldots, K$. Then we compute the Pearson correlation coefficient between the CE covariates at lag-one and the posterior means of the model residuals in each bin. CE covariates whose correlation with the residuals in one or more bin is significantly different from zero at the 5% level are selected for our model. This is a convenient yet simple method for variable selection. Though a more rigorous approach might instead make use of a shrinkage prior, our focus in this paper is sparsity in the matrix of autoregressive coefficients, as we discuss in Section 5. We find that $L = 5$ covariates are correlated with the residuals: nitrate, chemical oxygen demand (COD), ammonia, pH and phosphate. This selection of five covariates is supported by our exploratory data analysis. In Section 2.4, we found that several of the top 12 OTUs appeared to have a (contemporary) correlation with COD, ammonia and phosphate. We also found that one of the top 12 genera seemed to be correlated with nitrate and two of the top 12 classes had weak negative (contemporary) correlations with pH.

4.3. Missing data model

As discussed in Section 2.1, there are some missing OTU counts which lead to missing values in the scaled log counts of our bins y_{tj}, and these are treated as missing at random. In other words, we assume that the missingness is not related to the value of the count, allowing the missing data mechanism to be regarded as ignorable. As we are taking a Bayesian approach to inference, we simply treat these missing values as unobserved variables and marginalise the posterior over their values.

Some of the time series of CE covariates also contain missing observations. As with the missing y_{tj}, we treat these missing values as unobserved variables and average over our uncertainty in their values. This requires specification of a model for the (transformed) covariates X. We define $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tL})^T$ and allow the covariates to evolve according to a simple first-order autoregression

$$\boldsymbol{x}_t = \Phi_X \boldsymbol{x}_{t-1} + \tilde{\boldsymbol{\epsilon}}_t, \quad \tilde{\boldsymbol{\epsilon}}_t \sim \mathrm{N}_L\left(\mathbf{0}, \Sigma_X\right), \tag{7}$$

for $t = 2, \ldots, N$, where Φ_X is assumed to be diagonal, $\Phi_X = \mathrm{diag}(\phi_{X,1}, \ldots, \phi_{X,L})$.

In practice, we do not marginalise over the missing data analytically. Instead, their values are stochastically imputed, along with all other unknown parameters, in the Markov chain Monte Carlo algorithm that is used for computational inference; see Section 6.2.

5. Sparsity in the autoregressive coefficient matrix

In Section 4 we described the overall structure of our model. Now we consider the prior for the autoregressive coefficient matrix, which we allow to be sparse. We begin this section with a justification for allowing sparsity, before discussing different sparsity-inducing priors.

Time series of microbial data typically comprise high dimensional counts collected at a relatively small number of time points. Learning all pairwise interactions from such data can be prohibitively difficult and so models generally assume that most pairs of taxa do not interact (Ovaskainen et al., 2017). However, this is typically not motivated by biological reasoning. Instead, it is a pragmatic response to the challenges posed by the dimensionality of the data, which typically comprises measurements on a very large number of species (or groups of species) over a relatively small number of time points.

Learning about all the pairwise interactions in such a dynamic system is often prohibitively difficult. In the present case, the clustered OTU data comprise $N = 257$ observations on $K = 12$ bins and yet there are $K^2 = 144$ unknown coefficients in the autoregressive coefficient matrix A in (4). As such, an unstructured prior for the elements of A would yield very imprecise inferences. Further, for this particular application, given the circular time-ordering of the bins in the clustered OTU data, it is plausible that the previous abundances for all bins do not influence the current abundance of any particular bin once the abundances of neighbouring bins are known, though speculation was not tested in the paper. A sparse autoregressive matrix would reflect this notion, while providing a pragmatic solution to the challenge of dimensionality discussed above, by allowing the more influential coefficients to be learnt with a greater degree of precision. We also benefit from useful biological interpretation of the sparse matrix structure. Since the parameters in the autoregressive coefficient matrix can be interpreted as regression coefficients in a linear model, this is essentially a problem of variable selection which can be addressed by assigning a sparsity-inducing prior to the elements in the autoregressive coefficient matrix; see, for example, Gefang (2016) or Ahelegbey et al. (2016). A commonly adopted prior is a zero-mean scale-mixture of normals, in which the mixing distribution can either be discrete, as in spike-and-slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993), or continuous, as in the (regularised) horseshoe (Carvalho et al., 2009, 2010; Piironen and Vehtari, 2017). A simulation experiment, described in Section S4 of the Supplementary Materials, suggests that the horseshoe prior can yield very similar inferences to the spike-and-slab prior, but with greatly improved mixing during posterior sampling, and so we adopt a prior of this form for the parameters of the autoregressive coefficient matrix.

The horseshoe prior belongs to a class of global-local shrinkage priors in which the effect of a global hyperparameter is to encourage shrinkage of all coefficients towards zero. Local shrinkage hyperparameters, in one-to-one correspondence with the coefficients, then give the prior a heavy tail so as to retain support for large values of individual coefficients. It is widely known that the posterior can be very sensitive to the prior chosen for the global shrinkage parameter. Motivated by this observation, Piironen and Vehtari (2017) describe a principled methodology for the specification of this prior in the context of (generalised) linear regression. However, their attention was limited to univariate response variables. In order to apply similar ideas to the important choice of prior for the global shrinkage parameter in the vector autoregressive model, the remainder of this section describes an extension of the methodology to the class of linear regression models with a multivariate response, of which a vector autoregression can be regarded as a special case.

To define a hyperprior for the global shrinkage parameter, we first derive a general shrinkage factor matrix $\mathcal{K}_j$ in Section 5.1, which indicates how much a matrix of regression coefficients is shrunk towards its least squares estimator. This is achieved by considering the posterior distribution for the regression coefficients in a multivariate linear regression model conditional on the error variance matrix. From the shrinkage factor matrix, we define the effective number of non-zero coefficients m_{eff} (in Section 5.2), which is based on the assumption that shrinkage factor matrices tend to be diagonal binary matrices (Section S5.3 in the Supplementary Materials). By considering the prior expectation of m_{eff}, we then show how prior information on the number of non-zero coefficients can be incorporated into a hierarchical prior for the global shrinkage parameter that depends on the error variance matrix. Finally, in Section 5.3, we describe our final choice of prior for the autoregressive coefficient matrix.

5.1. The shrinkage factor matrix

Denote by $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iQ})^T$ a Q-variate response vector and by $\boldsymbol{x}_i$ a P-variate vector of covariates for experimental unit i. Under the multivariate linear regression model

$$\boldsymbol{y}_i = \mathrm{A}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}_Q(\mathbf{0}, \Sigma), \quad i = 1, \ldots, N, \tag{8}$$

where $\mathrm{A} = (a_{jk})$ is a $P \times Q$ matrix of regression coefficients and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iQ})^T$ is a vector of errors. We recover the order 1 vector autoregressive model when i indexes time, $\boldsymbol{x}_i = \boldsymbol{y}_{i-1}$ and $Q = P = K$. In matrix form, (8) can be written as

$$\mathrm{Y} = \mathrm{XA} + \mathrm{E} \tag{9}$$

where Y and X are $N \times Q$ and $N \times P$ data matrices with i-th rows $\boldsymbol{y}_i^T$ and $\boldsymbol{x}_i^T$, and E is a $N \times Q$ matrix of errors with i-th row $\boldsymbol{\epsilon}_i^T$.

Stacking the rows of Y and A into NQ- and PQ-vectors, respectively, let $\boldsymbol{y}^* = \mathrm{vec}(\mathrm{Y}^T) = (y_{11}, \ldots, y_{1Q}, y_{21}, \ldots, y_{NQ})^T$ and $\boldsymbol{a}^* = \mathrm{vec}(\mathrm{A}^T) = (a_{11}, \ldots, a_{1Q}, a_{21}, \ldots, a_{PQ})^T$. Now assume we give the regression coefficients $\boldsymbol{a}^*$ a horseshoe prior

$$\begin{aligned} a_{jk} | \lambda_{jk}, \tau &\sim \mathrm{N}(0, \tau^2 \lambda_{jk}^2), \\ \lambda_{jk} &\sim \mathrm{C}^+(0, 1), \end{aligned} \tag{10}$$

for $j = 1, \ldots, P$, $k = 1, \ldots, Q$, in which the λ_{jk} are the local shrinkage parameters, τ is the global shrinkage parameter and $\mathrm{C}^+(a, b)$ denotes a half Cauchy distribution with location a and scale b.

In Section S5.1 of the Supplementary Materials, we show that the conditional posterior for the regression coefficients $\boldsymbol{a}^*$ given the shrinkage parameters, $\Lambda^* = \mathrm{diag}(\lambda_{11}, \ldots, \lambda_{1Q}, \lambda_{21}, \ldots, \lambda_{PQ})^T$ and τ, along with the error variance Σ, is given by

$$\boldsymbol{a}^* | \Lambda^*, \tau, \Sigma, \boldsymbol{y}^* \sim \mathrm{N}_{PQ}(\boldsymbol{m}^*, \mathrm{V}^*) \tag{11}$$

where

$$\boldsymbol{m}^* = \tau^2 \Lambda^* \left[\tau^2 \Lambda^* + \left\{ \left(\mathrm{X}^T \mathrm{X} \right)^{-1} \otimes \Sigma \right\} \right]^{-1} \hat{\boldsymbol{a}}^*, \qquad \mathrm{V}^* = \left\{ \tau^{-2} \Lambda^{*-1} + \left(\mathrm{X}^T \mathrm{X} \otimes \Sigma^{-1} \right) \right\}^{-1}$$

in which

$$\hat{\boldsymbol{a}}^* = \mathrm{vec}(\hat{\mathrm{A}}^T) = \left\{ \left(\mathrm{X}^T \mathrm{X} \right)^{-1} \mathrm{X}^T \otimes \mathrm{I}_Q \right\} \boldsymbol{y}^*$$

and $\hat{\mathrm{A}}$ is the least squares estimator of A.

Suppose that the explanatory variables are uncorrelated with zero mean and variance $\mathrm{Var}(X_j) = s_j^2$ so that $\mathrm{X}^T \mathrm{X} \simeq N \mathrm{diag}(s_1^2, \ldots, s_P^2)$. Although in practice, it will rarely be the case that all covariates are uncorrelated, this pragmatic simplification is adopted here to facilitate theoretical development of an intuitive measure of sparsity. Setting $\mathrm{X}^T \mathrm{X} \simeq N \mathrm{diag}(s_1^2, \ldots, s_P^2)$ in the expression for the posterior mean $\boldsymbol{m}^*$ of $\boldsymbol{a}^*$ shows that

$$\begin{aligned} \tau^2 \Lambda^* \left[\tau^2 \Lambda^* + \left\{ \left(\mathrm{X}^T \mathrm{X} \right)^{-1} \otimes \Sigma \right\} \right]^{-1} &= \mathrm{blockdiag} \left\{ \tau^2 \Lambda_1 \left(\tau^2 \Lambda_1 + \frac{1}{N s_1^2} \Sigma \right)^{-1}, \ldots, \tau^2 \Lambda_P \left(\tau^2 \Lambda_P + \frac{1}{N s_P^2} \Sigma \right)^{-1} \right\} \\ &= \mathrm{blockdiag} \left(\mathrm{I}_Q - \mathcal{K}_1, \ldots, \mathrm{I}_Q - \mathcal{K}_P \right) \end{aligned}$$

where $\Lambda_j = \mathrm{diag}(\lambda_{j1}^2, \ldots, \lambda_{jQ}^2)$ is the j-th diagonal block of Λ^* and

$$\mathcal{K}_j = \left(\mathrm{I}_Q + N s_j^2 \tau^2 \Lambda_j \Sigma^{-1} \right)^{-1}, \quad j = 1, \ldots, P.$$

Hence we have

$$\boldsymbol{m}^* = \mathrm{blockdiag} \left(\mathrm{I}_Q - \mathcal{K}_1, \ldots, \mathrm{I}_Q - \mathcal{K}_P \right) \hat{\boldsymbol{a}}^* = \mathrm{vec}(\mathrm{M}^T)$$

where $\mathrm{M} = (m_{jk}) = \mathrm{E}(\mathrm{A} | \Lambda^*, \tau, \Sigma, \mathrm{Y})$. If we define $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jQ})^T$, $\hat{\boldsymbol{a}}_j = (\hat{a}_{j1}, \ldots, \hat{a}_{jQ})^T$ and $\boldsymbol{m}_j = (m_{j1}, \ldots, m_{jQ})^T$ for $j = 1, \ldots, P$ as the (transposed) columns of A, $\hat{\mathrm{A}}$ and M, respectively, then it is clear that

$$\boldsymbol{m}_j = \left(\mathrm{I}_Q - \mathcal{K}_j \right) \hat{\boldsymbol{a}}_j, \quad j = 1, \ldots, P$$

and so we can imagine constructing the posterior mean M^T of A^T column-wise; column j, corresponding to the coefficients of covariate j in the linear predictors of Y_1 through Y_Q, is a linear transformation of column j of the (transposed) least squares estimator $\hat{\mathrm{A}}^T$.

Since Λ_j and Σ are real and positive definite, the eigenvalues, $\eta_1, \ldots, \eta_Q$, of $N s_j^2 \tau^2 \Lambda_j \Sigma^{-1}$ must be real and positive. The eigenvalues of $\mathcal{K}_j^{-1} = \mathrm{I}_q + N s_j^2 \tau^2 \Lambda_j \Sigma^{-1}$ are therefore $1 + \eta_j > 1$ for $j = 1, \ldots, Q$, and hence the eigenvalues of $\mathcal{K}_j$, $1/(1 + \eta_j)$, must lie between 0 and 1, making it a convergent matrix. We can therefore regard $\mathcal{K}_j$, as the *shrinkage factor matrix* for coefficients $\boldsymbol{a}_j$ of covariate j. The size of the eigenvalues of $\mathcal{K}_j$ determine the extent to which the coefficients $\boldsymbol{a}_j$ are shrunk towards zero. Since the eigenvalues η_j are directly proportional to τ^2, as $\tau \to 0$, all eigenvalues of $\mathcal{K}_j$ approach 1 and we have $\mathcal{K}_j \to \mathrm{I}_Q$ and hence complete shrinkage. When $\tau \to \infty$ all eigenvalues of $\mathcal{K}_j$ approach 0 and we have $\mathcal{K}_j \to 0_Q$, where 0_Q denotes a matrix of zeros, and hence no shrinkage.

For an unstructured error variance matrix Σ, a closed form solution for the eigenvalues of the shrinkage factor matrix $\mathcal{K}_j$ is not available. However, it can be instructive to consider simpler parametric forms. For the vector autoregressive model in Section 4, we assume that Σ^{-1} is a symmetric, circulant, tridiagonal matrix, taking the form (3). In this case, letting $d_j^2 = N s_j^2 \tau^2 \sigma_0^{-2} \omega$ and $\tilde{\lambda}_{jk} = d_j \lambda_{jk}$ for $k = 1, \ldots, K$ we have

$$\mathcal{K}_j^{-1} = \begin{pmatrix} 1 + \omega^{-1} \tilde{\lambda}_{j1}^2 & \sigma_0^2 \tilde{\lambda}_{j1}^2 & 0 & 0 & \cdots & 0 & 0 & 0 & \sigma_0^2 \tilde{\lambda}_{j1}^2 \\ \sigma_0^2 \tilde{\lambda}_{j2}^2 & 1 + \omega^{-1} \tilde{\lambda}_{j2}^2 & \sigma_0^2 \tilde{\lambda}_{j2}^2 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \sigma_0^2 \tilde{\lambda}_{j3}^2 & 1 + \omega^{-1} \tilde{\lambda}_{j3}^2 & \sigma_0^2 \tilde{\lambda}_{j3}^2 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \sigma_0^2 \tilde{\lambda}_{j,K-1}^2 & 1 + \omega^{-1} \tilde{\lambda}_{j,K-1}^2 & \sigma_0^2 \tilde{\lambda}_{j,K-1}^2 \\ \sigma_0^2 \tilde{\lambda}_{jK}^2 & 0 & 0 & 0 & \cdots & 0 & 0 & \sigma_0^2 \tilde{\lambda}_{jK}^2 & 1 + \omega^{-1} \tilde{\lambda}_{jK}^2 \end{pmatrix}.$$

Although a closed-form solution for its eigenvalues and inverse are not available, in the special case when $\omega = 0$, so that $\Sigma = \sigma_0^{-2} \mathrm{I}_Q$, the shrinkage factor matrix reduces to

$$\mathcal{K}_j = \operatorname{diag}\left(\frac{1}{1+Ns_j^2\tau^2\lambda_{j1}^2\sigma_0^{-2}}, \ldots, \frac{1}{1+Ns_j^2\tau^2\lambda_{jQ}^2\sigma_0^{-2}}\right), \quad j=1,\ldots,P \tag{12}$$

so that each component of the posterior mean can be expressed as a product of a single shrinkage factor and the corresponding element of the least squares estimator

$$m_{jk} = \left(1 - \frac{1}{1+Ns_j^2\tau^2\lambda_{jk}^2\sigma_0^{-2}}\right)\hat{a}_{jk}, \quad j=1,\ldots,P,\ k=1,\ldots,Q\,.$$

Clearly the eigenvalues of $\mathcal{K}_j$ are simply its diagonal entries with a value near 0 or 1 indicating no shrinkage or complete shrinkage of the corresponding least squares estimate. The results for some other parametric forms are described in Section S5.2 of the Supplementary Materials.

5.2. Effective number of non-zero coefficients

In the case of a univariate response, there is a single shrinkage factor κ_j for each regression coefficient a_j. In the joint prior induced by the half-Cauchy densities for the local shrinkage parameters, each κ_j is conditionally independent of κ_k $(k \neq j)$ given the global shrinkage parameter τ and the error variance σ^2. The conditional priors for the κ_j can be derived in closed form and have u-shaped densities over the unit interval meaning that, *a priori*, most κ_j are either zero or one and so

$$m_{\text{eff}} = \sum_{j=1}^{P}(1-\kappa_j)$$

can be interpreted as the effective number of non-zero coefficients.

In the multivariate case, when the error variance Σ is diagonal, Section 5.1 showed that the shrinkage factor matrices $\mathcal{K}_j$ are also diagonal. It follows by direct analogy with the univariate result that the priors for the diagonal elements of $\mathcal{K}_j$ will be independent *a priori* with u-shaped densities. Therefore the same logic applies and we can interpret

$$m_{\text{eff}} = \sum_{j=1}^{P}\operatorname{tr}\left(\mathrm{I}_Q - \mathcal{K}_j\right) \tag{13}$$

as the effective number of non-zero coefficients. In the case of an unstructured error variance matrix Σ, conditional on Σ and τ, the mapping from Q-dimensional $(\lambda_{j1},\ldots,\lambda_{jQ})$ to Q^2-dimensional $\mathcal{K}_j$ is dimension-increasing, and so $\mathcal{K}_j$ must lie on a Q-dimensional manifold of $\mathbb{R}^{Q^2}$. As such, the independent, unit-median half-Cauchy distributions for the diagonal elements of Λ_j induce a joint distribution for $\mathcal{K}_j$ for which a density function does not exist. We can, nevertheless, explore the marginal and pairwise joint densities of the elements $\mathcal{K}_{j,k\ell}$ by simulation. For a range of values for the correlations in Σ, a simulation experiment described in Section S5.3 the Supplementary Materials reveals that the prior distribution for the elements of $\mathcal{K}_j$ assigns high probability to diagonal binary matrices. It does not, therefore, seem unreasonable to continue to interpret m_{eff}, as defined in (13), as the effective number of non-zero coefficients.

The prior expectation of the effective number of non-zero coefficients m_{eff}, conditional on τ and Σ is not generally available in closed form owing to the absence of a closed form expression for $\mathcal{K}_j$. However, in the special case when $\Sigma = \sigma^2\mathrm{I}_Q$, it follows immediately from the results in the univariate case that

$$\mathrm{E}_{\Lambda|\tau,\Sigma}(m_{\text{eff}}) = \frac{\sqrt{N}\tau\sigma^{-1}}{1+\sqrt{N}\tau\sigma^{-1}}PQ\,,$$

where we have assumed that the explanatory variables have been standardised to have variance equal to 1, that is, $s_j^2 = 1$ for $j=1,\ldots,P$. Suppose e_0 is our prior expectation for the number of non-zero coefficients in A. Then we can set $\mathrm{E}_{\Lambda|\tau,\Sigma}(m_{\text{eff}}) = e_0$, $\tau = \tau_0$ and solve for τ_0 to obtain

$$\tau_0 = \frac{e_0}{PQ - e_0}\frac{\sigma}{\sqrt{N}} \tag{14}$$

which demonstrates that τ should scale with $\sigma/\sqrt{N}$ if the prior expectation of the effective number of non-zero coefficients m_{eff} is to remain constant. We can then choose as our prior for τ

$$\tau|\sigma \sim \mathrm{C}^{+}(0,\tau_0^2) \tag{15}$$

which has conditional median equal to τ_0 given σ.

Of course, for general problems it may not be the case that we wish to restrict $\Sigma = \sigma^2 I_Q$; indeed this is not our choice for the vector autoregression in Section 4. In such cases, we can make our prior specification consistent with a central value in the prior for Σ by introducing a hyperparameter σ such that, say, $E(\Sigma|\sigma) = \sigma^2 I_Q$ or $E(\Sigma^{-1}|\sigma) = 1/\sigma^2 I_Q$. We can then construct our prior for (τ, Σ, σ) or $(\tau, \Sigma^{-1}, \sigma)$ hierarchically so that

$$\pi(\tau, \Sigma, \sigma) = \pi(\tau|\sigma)\pi(\Sigma|\sigma)\pi(\sigma)$$

or

$$\pi(\tau, \Sigma^{-1}, \sigma) = \pi(\tau|\sigma)\pi(\Sigma^{-1}|\sigma)\pi(\sigma) \tag{16}$$

with the conditional distribution for $\tau|\sigma$ specified in (15).

For the symmetric, circulant, tridiagonal precision matrix Σ^{-1} in (3) used in the vector autoregressive model, parameterised in terms of $\varpi_0 = \sqrt{2}(\sigma_0^{-2} + 2\omega)/2 > 0$ and $\varpi_1 = \sqrt{2}(\sigma_0^{-2} - 2\omega)/2 > 0$, we use a prior of the form (16) and require

$$E(\sigma_0^{-2}|\sigma) = \frac{E(\varpi_0|\sigma) + E(\varpi_1|\sigma)}{\sqrt{2}} = \frac{1}{\sigma^2} \quad \text{and} \quad E(\omega|\sigma) = \frac{E(\varpi_0|\sigma) - E(\varpi_1|\sigma)}{2\sqrt{2}} = 0$$

and hence

$$E(\varpi_0|\sigma) = E(\varpi_1|\sigma) = \frac{\sqrt{2}}{2\sigma^2}.$$

Specifically, for $i = 0, 1$, independently, we take

$$\varpi_i|\sigma \sim \text{Ga}\left(\frac{1}{c_\varpi^2}, \frac{\sqrt{2}\sigma^2}{c_\varpi^2}\right) \tag{17}$$

and then

$$\sigma \sim \text{LN}(m_\sigma, s_\sigma^2). \tag{18}$$

In the application in Section 7, we choose $c_\varpi = 1$ to maximise the conditional prior variance for ϖ_i given σ whilst keeping the density at zero finite. We then take $m_\sigma = 0$ and $s_\sigma = \sqrt{10}$ giving σ a distribution with median equal to one, consistent with our beliefs about the magnitude of the error variance. When Σ is unstructured, suitable choices for its prior are outlined in Section S6 of the Supplementary Materials.

5.3. Regularised horseshoe prior

The (original) horseshoe prior in (10) can pose some problems for the inferential scheme if one or more coefficient a_{jk} is only weakly identified by the likelihood. In such cases, the prior imparts little influence and the posterior remains long-tailed and difficult to sample. In order to address this problem, without compromising the interpretation of the effective number of non-zero coefficients, Piironen and Vehtari (2017) propose a modification to the horseshoe, called the regularised horseshoe, which would involve replacing (10) with

$$a_{jk}|\lambda_{jk}, \tau, c \sim \text{N}\left(0, \tau^2\tilde{\lambda}_{jk}^2\right), \qquad \tilde{\lambda}_{jk}^2 = \frac{c^2\lambda_{jk}^2}{c^2 + \tau^2\lambda_{jk}^2}, \qquad \lambda_{jk} \sim \text{C}^+(0, 1), \qquad \tau \sim \text{C}^+\left(0, \tau_0^2\right), \tag{19}$$

for $j = 1, \ldots, P$, $k = 1, \ldots, Q$. This can be regarded as a continuous analogue of a spike-and-slab prior, where the slab with infinite variance is replaced with a slab with finite variance c^2. The parameter $c > 0$ can be fixed or given a prior to reflect beliefs about the maximum possible size of the regression coefficients. We use precisely a prior of this form for the parameters in the autoregressive coefficient matrix A in our VAR(1) model. Though we take $e_0 = E_{\Lambda|\tau,\Sigma}(m_{\text{eff}}) = 12$ to reflect our beliefs that only the $K = 12$ diagonal elements of A are likely to be non-zero, we found the posterior to be largely insensitive to plausible changes to this value; see Section S7 of the Supplementary Materials. We then choose

$$c^2 \sim \text{IG}(2, 8) \tag{20}$$

to reflect the idea that, after scaling the data so that each bin has standard deviation roughly equal to 1, we would not expect any regression coefficient to exceed around 5 in absolute value. Note that $\text{IG}(g, h)$ denotes an inverse gamma distribution with shape g and scale h.

6. Bayesian inference

In this section we begin by completing our description of the prior distribution for the model parameters. We then describe our approach to sampling from the posterior distribution using Hamiltonian Monte Carlo, implemented through Stan.

6.1. Prior distribution

There are several parameters in the hierarchical VAR(1) model. First, there is the matrix of autoregressive coefficients A with its global shrinkage parameter $\tau \in \mathbb{R}_+$, the matrix of local shrinkage parameters Λ^* and the regularising parameter c^2. Additionally, there are the parameters for the precision matrix Σ^{-1} of the errors, denoted by $\boldsymbol{\varpi} = \{\varpi_0, \varpi_1\} \in \mathbb{R}_+^2$ and the parameter σ in their hierarchical prior. There are the harmonic regression coefficients for $\boldsymbol{\mu}_t$, denoted by $\boldsymbol{\theta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\}$. We also have the coefficients of the CE covariates X and the intercepts in B, with unknown means $\boldsymbol{\mu}_B = (\mu_{B_0}, \dots, \mu_{B_L})$ and variances $\boldsymbol{\sigma}_B = (\sigma^2_{B_0}, \dots, \sigma^2_{B_L})$ for the $L+1$ rows of B. Finally, we have the parameters in $\Phi_X = \mathrm{diag}(\phi_{X,1}, \dots, \phi_{X,5})$ and Σ_X in the missing data model.

We adopt a prior distribution in which these various parameter blocks are independent, with hierarchical structure within blocks:

$$\begin{aligned}\pi\left(\mathrm{A}, \tau, \Lambda^*, c^2, \boldsymbol{\varpi}, \sigma, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B, \Phi_X, \Sigma_X\right) = {} & \pi\left(\mathrm{A}|\Lambda^*, \tau, c^2\right)\pi\left(\tau|\sigma\right)\pi(\sigma)\pi(\Lambda^*)\pi(c^2)\pi\left(\boldsymbol{\varpi}|\sigma\right) \\ & \times \pi\left(\boldsymbol{\theta}\right)\pi\left(\mathrm{B}|\boldsymbol{\mu}_B, \boldsymbol{\sigma}_B\right)\pi\left(\boldsymbol{\mu}_B\right)\pi\left(\boldsymbol{\sigma}_B\right) \\ & \times \pi\left(\Phi_X\right)\pi\left(\Sigma_X\right).\end{aligned} \tag{21}$$

Our main focus in this paper has been the joint distribution encoded through the right-hand-side of the first line of (21). The distributions $\pi\left(\mathrm{A}|\Lambda^*, \tau, c^2\right)$, $\pi\left(\tau|\sigma\right)$ and $\pi(\Lambda^*)$ are given in (19); $\pi(c^2)$ in (20); $\pi\left(\boldsymbol{\varpi}|\sigma\right) = \pi(\varpi_0|\sigma)\pi(\varpi_1|\sigma)$ and $\pi(\sigma)$ in (17) and (18), respectively.

For the parameters in the harmonic regression component of the time-varying mean, we take $\boldsymbol{\beta}_j \sim \mathrm{N}_K(\mathbf{0}, \mathrm{V}_\beta)$ and $\boldsymbol{\gamma}_j \sim \mathrm{N}_K(\mathbf{0}, \mathrm{V}_\gamma)$ independently for $j = 1, 2$, with $\mathrm{V}_\beta = \mathrm{V}_\gamma = 100\mathrm{I}_K$. For the coefficients of the CE covariates, we adopt hierarchical priors for the $b_{\ell k}$, such that, for each $\ell = 0, \dots, L$ independently, we have $b_{\ell k}|\mu_{B_\ell}, \sigma^2_{B_\ell} \sim \mathrm{N}\left(\mu_{B_\ell}, \sigma^2_{B_\ell}\right)$ independently for $k = 1, \dots, K$, $\mu_{B_\ell} \sim \mathrm{N}\left(a_\alpha, b_\alpha^2\right)$ and $\sigma^2_{B_\ell} \sim \mathrm{IG}\left(c_\alpha, d_\alpha\right)$. To give $\mathrm{E}(b_{\ell k}) = 0$, $\mathrm{Corr}\left(b_{\ell j}, b_{\ell k}\right) = 0.95$ and $\mathrm{Var}\left(b_{\ell j}\right) = 100$, we select $a_\alpha = 0$, $b_\alpha = \sqrt{95}$, $c_\alpha = 2.25$ and $d_\alpha = 6.25$. This allows borrowing of strength across bins between the coefficients of each CE covariate.

Finally, for the parameters in the missing data model, we adopt the prior $\phi_{X,\ell} \sim \mathrm{Beta}\left(a_\phi, b_\phi\right)$ independently for $\ell = 1, \dots, 5$ and $\Sigma_X \sim \mathcal{W}^{-1}\left(\mathrm{S}_X, \nu_X\right)$, with $a_\phi = b_\phi = 2$, $\mathrm{S}_X = \mathrm{I}_5$ and $\nu_X = 9$. Note that $\mathcal{W}^{-1}\left(\Psi, \nu\right)$ denotes an inverse Wishart distribution with scale matrix Ψ and ν degrees of freedom.

6.2. Posterior distribution and computation

Combining the prior (21) with the first-order Markovian likelihood (4) and the missing data model (7) using Bayes theorem yields the posterior distribution as

$$\begin{aligned}\pi\left(\mathrm{A}, \tau, \Lambda^*, c^2, \boldsymbol{\varpi}, \sigma, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B, \Phi_X, \Sigma_X|\mathrm{Y}, \mathrm{X}\right) \propto {} & \pi\left(\mathrm{A}, \tau, \Lambda^*, c^2, \boldsymbol{\varpi}, \sigma, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B, \Phi_X, \Sigma_X\right) \\ & \times \pi\left(\mathrm{Y}|\mathrm{A}, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\varpi}, \mathrm{X}\right)\pi\left(\mathrm{X}|\Phi_X, \Sigma_X\right),\end{aligned}$$

where the likelihood can be written as

$$\pi\left(\mathrm{Y}|\mathrm{A}, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\varpi}, \mathrm{X}\right) = \prod_{t=2}^{N} \pi\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \mathrm{A}, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\varpi}, \boldsymbol{x}_{t-1}\right)$$

with $\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \mathrm{A}, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\varpi}, \boldsymbol{x}_{t-1} \sim \mathrm{N}_K\left(\boldsymbol{\mu}_t + \mathrm{A}\left(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_{t-1}\right), \Sigma\right)$. Here we have implicitly conditioned on the observed value of $\boldsymbol{y}_1$ so that the contribution of the marginal model $\pi\left(\boldsymbol{y}_1|\mathrm{A}, \boldsymbol{\theta}, \mathrm{B}, \boldsymbol{\varpi}, \boldsymbol{x}_0\right)$ can be ignored. This is reasonable as we have a large enough sample size that little information will be lost in doing so. The missing data model $\pi(\mathrm{X}|\Phi_X, \Sigma_X)$ is constructed in an analogous fashion.

This posterior distribution is analytically intractable, so we resort to sampling from it using Markov chain Monte Carlo methods. More specifically, we use Hamiltonian Monte Carlo (HMC) (Neal, 2011; Girolami and Calderhead, 2011) which is a gradient-based, auxiliary variable method which is well-suited for inference in hierarchical models (Betancourt and Girolami, 2015). The HMC algorithm was implemented using `rstan` (Stan Development Team, 2020), the R interface to the Stan software (Carpenter et al., 2017). Stan requires users to write a program in the probabilistic Stan modelling language, the role of which is to provide instructions for computing the logarithm of the kernel of the posterior density function. The

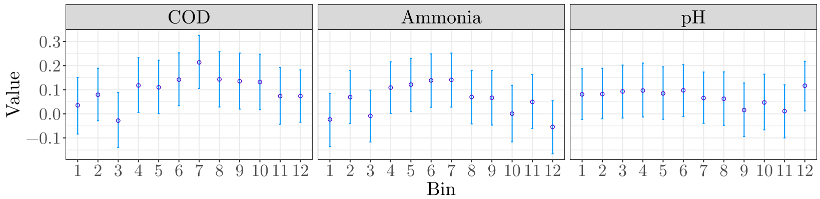

Fig. 4. Posterior means (∘) and 95% credible intervals (−) for COD, ammonia and pH.

Stan software then automatically tunes and runs a Markov chain simulation to sample from the resulting posterior. The Stan program used in our application can be found in Section S8 of the Supplementary Materials.

We ran the algorithm for 10K iterations, with a warm-up period of 3000 iterations. In the interests of saving memory, the output is thinned to leave us with 1000 samples from the posterior. The usual graphical and numerical diagnostic checks gave no evidence of any lack of convergence and mixing was good. We present the results in the next section based on these posterior samples.

7. Application to WWTP data

Now we discuss our findings after fitting the model to the data. We look at the posterior means and 95% credible intervals (CIs) of the model parameters. If zero is contained in a CI, we use this as a discriminator to suggest that the parameter's value may be (close to) zero. However, we emphasise that this does not necessarily mean that there is not considerable support for a positive or negative coefficient.

7.1. Time varying mean

We begin with the time varying mean $\boldsymbol{\mu}_t$ of the model, which is described by two seasonal harmonics and time-varying CE covariates. Recall that $\boldsymbol{\mu}_t = \boldsymbol{\alpha}_t + \sum_{j=1}^{2} \boldsymbol{\beta}_j \sin(2\pi jt/52) + \sum_{j=1}^{2} \boldsymbol{\gamma}_j \cos(2\pi jt/52)$, where $\boldsymbol{\alpha}_t$ has k-th element $\alpha_{tk} = b_{0k} + b_{1k}x_{t-1,1} + \cdots + b_{Lk}x_{t-1,L}$ and $\boldsymbol{X}_t$ are the transformed measurements of our chosen CE covariates at time t. We first examine the posterior means and CIs of the regression coefficients B for the CE covariates and provide possible biological explanations for the results. Then we look at the results for the harmonic regression coefficients $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ for $j = 1, 2$.

7.1.1. Chemical and environmental covariates

There are five covariates in the model: nitrate, COD, ammonia, pH and phosphate. For all 12 bins, the CIs for nitrate and phosphate shown in Supplementary Fig. S27 all include zero, suggesting that neither of these covariates has a linear relationship with the time varying mean of any of the bins, although for phosphate we note that the CIs for bins 2 and 3 only just overlap zero. Despite all of the CIs overlapping zero, there is a clear pattern for phosphate, with "winter blooming" bins showing a positive relationship and "summer blooming" bins showing a negative relationship, though this might be attributable to the nature of the autocorrelation between bins.

From Fig. 4, bins 4 to 10 have positive regression coefficients with COD, with bin 7 having the largest regression coefficient with a posterior mean of 0.2136 (4 d.p.). Ammonia has a positive regression coefficient with four bins (4 to 7) and, as we saw with COD, bin 7 has the largest regression coefficient with a posterior mean of 0.1412 (4 d.p.). Finally, bin 12 seems to have a positive relationship with pH, with its regression coefficient having a posterior mean of 0.1167 (4 d.p.).

Removal of ammonia and other pollutants is essential in the treatment of wastewater and ammonia is removed through nitrification by bacteria. In the nitrogen cycle, nitrification is a two-step process of ammonia oxidation then nitrite oxidation. Bacteria from the genus *Nitrosomonas* can oxidise ammonia to nitrite (Wetzel, 2001), although there are other ammonia oxidising microorganisms (AOM) too. *Nitrobacter* bacteria from the same phylum as *Nitrosomonas* oxidise nitrite to nitrate but are difficult to detect *in-situ*. Wagner et al. (1996) suggested that this could be because they have a minor role in WWTPs and although Alawi et al. (2009) agreed that their role is small, they also noted that lack of detection does not necessarily mean lack of presence. In the AS, no *Nitrobacter* counts are recorded. This could suggest that in our WWTP other nitrite oxidising bacteria (NOB), for example, *Nitrospira* and *Ca. Nitrotoga*, are responsible for nitrite removal or that the *Nitrobacter* bacteria simply have not been detected, as seen in the literature.

Until recently, *Nitrospira* were considered solely NOB (Mehrani et al., 2020). Daims et al. (2015) and van Kessel et al. (2015) independently discovered a single microorganism from the genus *Nitrospira* that can carry out complete nitrification through the *comammox* (complete oxidation of ammonia to nitrate) process. Additionally, it has been found that there is a reciprocal feeding interaction between nitrifiers. Some species of *Nitrospira* are able to convert urea to ammonia and carbon dioxide, which means they can supply AOM with ammonia and in return receive nitrite produced by ammonia oxidation (Koch et al., 2015). An OTU from the genus *Nitrospira* is one of the most abundant OTUs within bin 4 and this OTU could be capable of comammox which could provide a reasonable explanation as to why there is a positive coefficient for ammonia and bin 4. Furthermore, most microorganisms need ammonia to grow via nitrogen assimilation, so that might explain why we see a positive relationship between bins 4 to 7 and ammonia.

To understand why some of the other bins may have a relationship with COD, we look at the most abundant OTUs within some of the bins. Supplementary Table S5 shows the genera of the top six OTUs in each bin. An OTU from the genus *Terrimonas* is the most abundant in bin 4, with a median within-bin relative abundance of around 22.8%. Bacteria from this genus assimilate organic compounds such as sugars and proteins (McIlroy et al., 2015). This provides a possible explanation as to why bin 4 has a positive relationship with COD.

Of the 1274 OTUs in bin 5, the most abundant OTU based on median within-bin relative abundance ($\sim 9.1\%$) is from the genus *Zoogloea*. Bacteria from this genus are highly active oxidisers of organic compounds (Dugan, 1981). Recalling that covariates are incorporated into the model via lag-one regression, the transformed COD measurement from the previous time point is used to model the intercept of the time varying mean at the current time point. If COD is high then this would suggest that there is a larger amount of organic compounds available for the *Zoogloea* bacteria to oxidise for energy and grow, thus explaining the positive coefficient between the bin containing *Zoogloea* and COD. However, this could result in the amount of organic compounds (and COD) decreasing which in turn could eventually slow the growth rate of the *Zoogloea* bacteria. More organic compounds can migrate into the system as more wastewater enters the WWTP which could then cause the COD to rise again. This describes a predator-prey-like dynamic and demonstrates that the relationships between the covariates and bins (of OTUs) are unlikely to be simple.

An OTU from the genus *Leptothrix* is the most abundant OTU in bin 7 based on median within-bin relative abundance ($\sim 10.4\%$). Species from this genus typically oxidise iron and manganese (McIlroy et al., 2015). The second most abundant OTU is from the genus *Dechloromonas* with a median within-bin relative abundance of around 10%. In Section 2.4, we saw that some of the top genera were correlated with COD (Fig. 2a), where *Dechloromonas* had a fairly weak positive correlation and *Leptothrix* did not appear in the top 12 genera in the AS tank. Some species of *Dechloromonas* are *polyphosphate-accumulating organisms* (PAOs) and some species have a role in *denitrification* (McIlroy et al., 2015). PAOs are bacteria that aid the removal of organic compounds containing phosphorus from wastewater. Denitrification is the reduction of nitrate to the eventual product of nitrogen gas, following a series of intermediate gaseous nitrogen oxide products. Nitrate and phosphorus both contribute to the COD of wastewater. Applying logic similar to that discussed for the *Zoogloea* bacteria in bin 5, a positive and likely non-linear relationship between COD and bin 7 seems sensible.

Finally, we focus on bin 12, which is the only bin that has a non-zero (positive) coefficient with pH. OTUs 15, 33 and 65, from the genus *Rhodobacter*, represent about 23.9% of bin 12 on average. Most *Rhodobacter* strains grow at an optimal pH range of 6.5 - 7.5 (Imhoff, 2015). The pH ranges from 5.02 to 7.5 with a median of 6.53, thus providing a possible explanation as to why bin 12 has a positive relationship with pH. Fig. 2a in Section 2.4 does not seem to indicate a correlation between *Rhodobacter* and pH. However, OTU 15 possibly has a weak positive correlation with pH (Fig. 1a). It is also important to remember that the heatmaps show correlations, not lag-one correlations. Calculating both the correlation (0.1155) and lag-one correlation (0.1679) between pH and *Rhodobacter*, we see that the lag-one correlation is stronger, thus corroborating our results. Furthermore, this relationship remains after allowing for other covariates and interactions, which highlights the benefit of the model; this relationship may otherwise go unnoticed.

7.1.2. Harmonic regression coefficients

Now we look at the harmonic regression coefficients of the model. Supplementary Figs. S28 and S29 show the posterior means and 95% CIs for the harmonic regression coefficients $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ for $j = 1, 2$. The change in the values of the β_{jk} and γ_{jk} across bins, $k = 1, \ldots, 12$, for the first harmonic ($j = 1$) can be explained by our chosen clustering method, which is based on the idea that the OTUs display seasonal variation and peak in different months. Recalling from Section 3.1, OTUs in bin 1 peak in February, OTUs in bin 2 peak in January, OTUs in bin 3 peak in December and so on. Based on the CIs, it would seem that only bins 2, 3 and 4 seem to have non-zero coefficients for the second harmonic, suggesting that their scaled log counts do not follow a pattern as simple as a sinusoid. Recall that in the time series plots of the bins in Fig. 3 (Section 3.1), we saw that the annual peaks were not as obvious in bins 2 and 3, suggesting a sinusoid may not be such a good descriptor. Perhaps, this is why we have non-zero coefficients for the second harmonics for these two bins. Reviewing the time series plots again, we can also see that in bin 4, there seem to be two peaks within 2013, with a smaller peak in the middle of the year and a larger peak around October. This might explain why we have a non-zero coefficient for the second harmonic in bin 4. Supplementary Fig. S30 shows posterior means of the time varying means $\boldsymbol{\mu}_t$ and the 95% CIs plotted over the scaled log counts for each bin. The seasonal patterns of each bin seem to have been captured fairly well.

7.2. Matrix of autoregressive coefficients

A vector autoregression is stationary if and only if the roots of the characteristic equation lie outside the unit circle; for models of order one, this means stationarity occurs if and only if the spectral radius of the single autoregressive coefficient matrix A is less than one. Interestingly, all the posterior samples of A in our analysis of the WWTP data lie within this stationary region. This suggests that the groups of microbes represented by our twelve bins are in a stable state after allowing for seasonality and the effects of the CE covariates. This is consistent with recent findings on the microbial community structure in the AS of WWTPs (Shchegolkova et al., 2016).

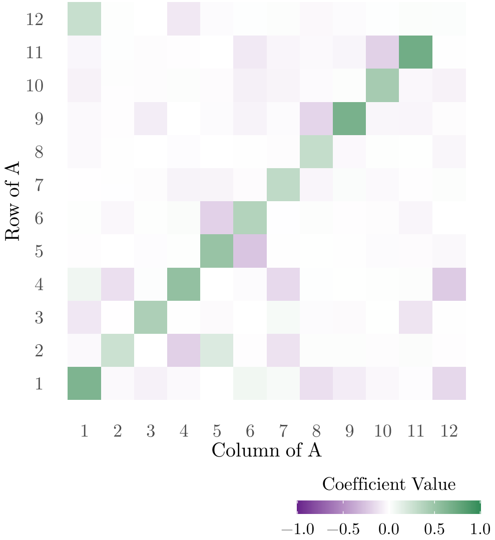

Fig. 5. Heatmap of the posterior means of the autoregressive coefficients. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

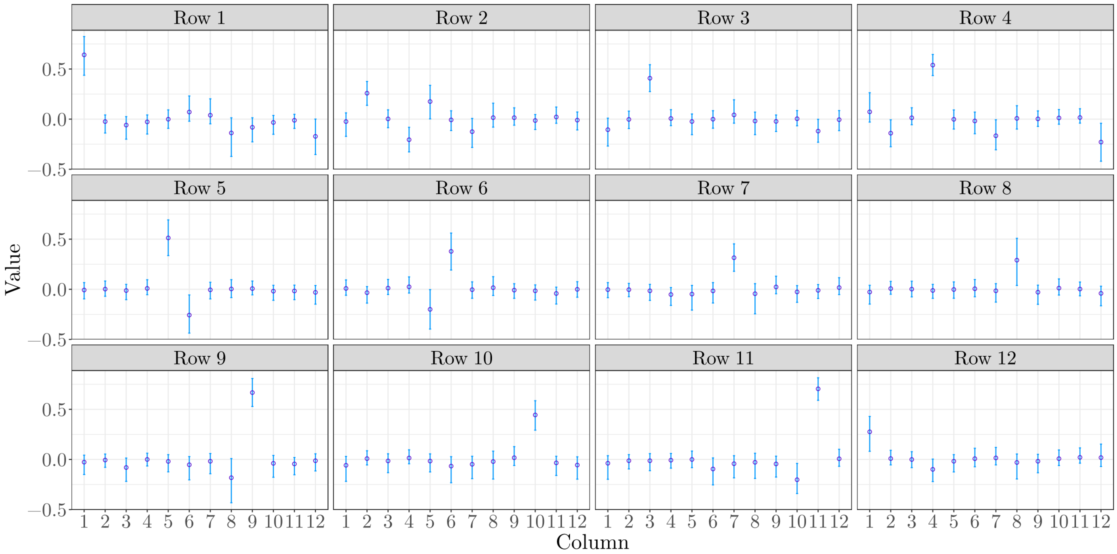

Fig. 6. Posterior means (∘) and 95% credible intervals (—) of the autoregressive coefficients.

The matrix of autoregressive coefficients is also informative about the relationships between bins. The posterior means of the autoregressive coefficients are shown in a heatmap in Fig. 5 and they are also shown in Fig. 6 with their corresponding 95% CIs. From the heatmap, we can see that the matrix of autoregressive coefficients based on posterior means is fairly sparse. With the exception of bin 12, all the bins have a positive autoregressive coefficient with themselves. In other words, the scaled log count of the previous time point seems to have a positive relationship with the scaled log count at the current time point, which seems sensible. Bins 1, 4, 5, 9 and 11 have particularly large "within-bin" autoregressive coefficients with posterior means larger than 0.5. It is surprising that the $a_{12,12}$ is a near-zero coefficient, with a posterior mean of 0.019 (3 d.p.). It could be that $y_{t,12}$ is better explained by $y_{t-1,1}$ than $y_{t-1,12}$. Bin 1 peaks in February and bin 12 peaks in March and the posterior mean of $a_{12,1}$ is positive (0.276), so this does not seem unreasonable. The notably larger diagonal values in A suggest that a possible improvement to the regularised horseshoe in application to vector autoregressions might allow the diagonal and off-diagonal elements to have their own global shrinkage parameters.

In addition to the within-bin autoregressive coefficients a_{kk}, we see from the CIs in Fig. 6 that there is evidence for a few non-zero "between-bin" posterior autoregressive coefficients a_{jk}, $j \neq k$. The posterior means for these coefficients whose CIs do not overlap zero are listed in Table 2. Apart from $a_{12,1}$, the non-zero between-bin coefficients are smaller than all the non-zero within-bin coefficients.

Table 2
Posterior means (3 d.p.) of the non-zero between-bin coefficients.

Coefficient	Posterior Mean	Coefficient	Posterior Mean
$a_{2,4}$	-0.204	$a_{4,12}$	-0.229
$a_{2,5}$	0.176	$a_{5,6}$	-0.257
$a_{3,11}$	-0.119	$a_{6,5}$	-0.200
$a_{4,2}$	-0.14	$a_{11,10}$	-0.202
$a_{4,7}$	-0.166	$a_{12,1}$	0.276

The non-zero between-bin coefficients constitute novel biological findings. To aid interpretation we look at the most abundant OTUs in each bin again. As mentioned above, the most abundant OTU in bin 5 is from the genus *Zoogloea*. The second most abundant OTU is from the genus *Acidovorax*. In bin 6, the second most abundant OTU is from the genus *Dechloromonas*, which as mentioned above is capable of nitrite reduction, as well as sulphate reduction. This is also true for *Zoogloea* and *Acidovorax* bacteria (McIlroy et al., 2015). Perhaps these bacteria amongst others that are not in the most abundant OTUs are competing for resources such as nitrite and sulphate, resulting in the negative autoregressive coefficients.

As stated above, an OTU from the strictly aerobic genus *Terrimonas* is the most abundant in bin 4. The second most abundant OTU is from the genus *Ca. Microthrix*, which is also described as aerobic in McIlroy et al. (2015). The top two OTUs in bin 2 are from the family *Rhodobacteraceae* with unknown genera. There are at least 288 known species from 99 genera (Pujalte et al., 2014) in the family *Rhodobacteraceae*, any of which the top two OTUs could be from. However, the third most abundant OTU, representing on average 14.5% of bin 2, is from the genus *Haematobacter* from the same family, which are aerobic bacteria. We cannot determine the genera of the top two OTUs but they may be aerobic, especially as most *Rhodobacteraceae* are aerobic (Pujalte et al., 2014). Perhaps there are negative interactions between bins 2 and 4 because aerobic microorganisms in both bins are competing for oxygen.

7.3. *Precision matrix for errors*

Recall that the errors in our model $\boldsymbol{\epsilon}_t$ follow a $\mathrm{N}_K(0, \Sigma)$ distribution and we have a symmetric, tridiagonal, circulant precision matrix for the errors, shown in (3). The posterior means for ω_0 and ω_1 are 6.7354 and -3.2183 (to 4 d.p.) respectively, with standard deviations 0.1926 and 0.0987 (to 4 d.p.). The covariance matrix for the errors Σ is a symmetric, circulant matrix. The correlation matrix associated with Σ is therefore defined by the lag-k correlations ρ_k for $k = 1, \dots, 6$. Supplementary Fig. S31 shows the posterior means and 95% CIs for $\rho_1, \dots, \rho_6$. All of the CIs lie above zero which provides evidence of between-bin correlation in the errors.

8. Discussion

The main objectives of this paper were to assess whether a model based on a VAR(1) process could provide useful biological insights into the microbial interactions in the AS of a WWTP and whether there was any evidence of chemical and environmental effects. Microorganisms in AS are responsible for biologically treating wastewater. Gaining an understanding of the complex network of microbial interactions is important to ensure a WWTP can continue functioning or, better still, be improved (Cydzik-Kwiatkowska and Zielińska, 2016). As is commonly found in metagenomics studies, our data suffer from high-dimensionality and sparsity. Owing to the evidence of seasonality in the data, we chose a seasonal phase-based clustering approach to address both issues.

Often, in time series metagenomics, gLV differential equations are used to model non-linear dynamics of the microbial communities of interest. However, we chose a more parsimonious option, which allows explicit modelling of the error, and developed a Bayesian hierarchical VAR(1) model for our clustered data. This is a simple first-order approximation to a gLV model. The circular time-ordering of the bins suggested a sparse autoregressive coefficient matrix would be sensible. We used a regularised horseshoe prior to allow for this. The posterior can be very sensitive to the choice of prior for the global shrinkage parameter in the horseshoe prior. We therefore extended the work of Piironen and Vehtari (2017), who considered its choice in the context of linear regression for a univariate response, to the multivariate setting. This gives a principled methodology for constructing the prior based on prior beliefs about the degree of sparsity. We gave the errors of our model a symmetric, circulant, tridiagonal precision matrix to complement the chosen clustering method. To capture the seasonal variation in each bin, we used a harmonic regression to fit a time varying mean, in which the CE data were incorporated.

The approach and modelling described in this paper is not without limitations. First, clustering the OTUs into bins prior to modelling complicated the interpretations of microbial interactions and the chemical and environmental effects because the bins do not necessarily represent OTUs with identical behaviours. Future work will therefore investigate clustering the data into a much larger number of bins, each containing a smaller number of OTUs, which will then be modelled as a time-series of multinomial counts. These ideas have already been applied in the simpler setting where OTU count vectors are treated as independent and identically distributed. For example, Zhang and Lin (2019) uses a logistic normal multinomial model with a sparse correlation structure to model counts of 87 genera from gut microbiome data collected from 98 subjects. Second, notwithstanding the clustering issue, the VAR(1) model used in this analysis is only a simplification of the biologically-motivated stochastic Lotka-Volterra model. Indeed, as explained in Section 1.1, linearisation results in omission of the second-order non-linear interactions. As such, future work will explore methods for fitting stochastic Lotka-Volterra models (Xu et al., 2020) and compare inferences to those obtained under the VAR(1) model.

Despite these drawbacks, we were able to achieve both of the objectives of the study. After fitting the model to our WWTP data, we identified possible relationships amongst bins and between bins and CE covariates by inspecting the posterior distributions obtained for the parameters in the model. For example, we found that some bins seem to have a positive relationship with COD and ammonia. After looking at the most abundant genera in each bin and biological literature, we also found evidence to suggest that microorganisms may be competing for resources. Altogether, the analysis provides an interesting insight into the dynamics of the microbial communities present in the AS of the WWTP.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), Centre for Doctoral Training in Cloud Computing for Big Data (grant number EP/L015358/1). This work was also supported by the EPSRC (grant number EP/N510129/1) via the Alan Turing Institute project “Streaming data modelling for real-time monitoring and forecasting”.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107659.

References

Ahelegbey, D.F., Billio, M., Casarin, R., 2016. Bayesian graphical models for structural vector autoregressive processes. J. Appl. Econom. 31, 357–386.

Alawi, M., Off, S., Kaya, M., Spieck, E., 2009. Temperature influences the population structure of nitrite-oxidizing bacteria in activated sludge. Environ. Microbiol. Rep. 1, 184–190.

Betancourt, M., Girolami, M., 2015. Hamiltonian Monte Carlo for hierarchical models. In: Upadhyay, S.K., Singh, U., Dey, D.K., Loganathan, A. (Eds.), Current Trends in Bayesian Methodology with Applications, first ed. CRC Press, Boca Raton, FL, pp. 79–101. Chapter 4.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., et al., 2016. MDSINE: microbial dynamical systems inference engine for microbiome time-series analyses. Genome Biol. 17, 121.

Bunge, J., Willis, A., Walsh, F., 2014. Estimating the number of species in microbial diversity studies. Annu. Rev. Stat. Appl. 1, 427–445.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. J. Stat. Softw. 76, 1–32.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009. Handling sparsity via the horseshoe. J. Mach. Learn. Res. 5, 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Curtis, T.P., Head, I.M., Graham, D.W., 2003. Peer reviewed: theoretical ecology for engineering biology. Environ. Sci. Technol. 37, 64–70.

Cydzik-Kwiatkowska, A., Zielińska, M., 2016. Bacterial communities in full-scale wastewater treatment systems. World J. Microbiol. Biotechnol. 32, 66.

Daims, H., Lebedeva, E.V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., et al., 2015. Complete nitrification by Nitrospira bacteria. Nature 528, 504–509.

Dam, P., Fonseca, L.L., Konstantinidis, K.T., Voit, E.O., 2016. Dynamic models of the complex microbial metapopulation of Lake Mendota. npj Syst. Biol. Appl. 2, 16007.

David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., Alm, E.J., 2014. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 15, R89.

Dugan, P.R., 1981. The genus Zoogloea. In: Starr, M.P., Stolp, H., Trüper, H.G., Balows, A., Schlegel, H.G. (Eds.), The Prokaryotes: A Handbook on Habitats, Isolation, and Identification of Bacteria. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 764–770.

Eiler, A., Heinrich, F., Bertilsson, S., 2012. Coherent dynamics and association networks among lake bacterioplankton taxa. ISME J. 6, 330–342.

Faust, K., Lahti, L., Gonze, D., de Vos, W.M., Raes, J., 2015. Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr. Opin. Microbiol. 25, 56–66.

Fisher, C.K., Mehta, P., 2014. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS ONE 9, 1–10. https://doi.org/10.1371/journal.pone.0102451.

Gefang, D., 2016. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. Int. J. Forecast. 30, 1–11.

Gelfand, A.E., Fuentes, M., Guttorp, P., Diggle, P., 2010. Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88, 881–889.

Gibbons, S.M., Kearney, S.M., Smillie, C.S., Alm, E.J., 2017. Two dynamic regimes in the human gut microbiome. PLoS Comput. Biol. 13, 1–20.

Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Stat. Soc. B 73, 123–214.

Goyal, A., Dubinkina, V., Maslov, S., 2018. Multiple stable states in microbial communities explained by the stable marriage problem. ISME J. 12, 2823–2834.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Imhoff, J.F., 2015. Rhodobaca. In: Bergey's Manual of Systematics of Archaea and Bacteria. American Cancer Society, pp. 1–4.

Kaul, A., Mandal, S., Davidov, O., 2017. Analysis of microbiome data in the presence of excess zeros. Front. Microbiol. 8, 2114.

van Kessel, M.A.H.J., Speth, D.R., Albertsen, M., Nielsen, P.H., Op den Camp, H.J.M., Kartal, B., Jetten, M.S.M., Lücker, S., 2015. Complete nitrification by a single microorganism. Nature 528, 555–559.

Koch, H., Lücker, S., Albertsen, M., Kitzinger, K., Herbold, C., Spieck, E., et al., 2015. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. Proc. Natl. Acad. Sci. 112, 11371–11376.

Konopka, A., Lindemann, S., Fredrickson, J., 2015. Dynamics in microbial communities: unraveling mechanisms to identify principles. ISME J. 9, 1488–1495.

Lee, K.H., Coull, B., Moscicki, A.B., Paster, B., Starr, J., 2018. Bayesian variable selection for multivariate zero-inflated models: application to microbiome count data. Biostat.

Lotka, A.J., 1926. Elements of physical biology. Sci. Prog. Twent. Century (1919–1933) 21, 341–343.

McIlroy, S.J., Saunders, A.M., Albertsen, M., Nierychlo, M., McIlroy, B., Hansen, A.A., et al., 2015. MiDAS: the field guide to the microbes of activated sludge. Database 2015.

Mehrani, M., Sobotka, D., Kowal, P., Ciesielski, S., Makinia, J., 2020. The occurrence and role of Nitrospira in nitrogen removal systems. Bioresour. Technol. 303, 122936.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Am. Stat. Assoc. 832, 1023–1032.

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A., Hélias, A., Irlinger, F., 2008. Microbial interactions within a cheese microbial community. Appl. Environ. Microbiol. 74, 172–181.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.L. (Eds.), Handbook of Markov Chain Monte Carlo. In: Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pp. 113–162.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B., Abrego, N., 2017. How are species interactions structured in species-rich communities? A new method for analysing time-series data. Proc. - Royal Soc. B, Biol. Sci. 284, 20170768.

Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkages priors. Electron. J. Stat. 11, 5018–5051.

Pujalte, M.J., Lucena, T., Ruvira, M.A., Arahal, D.R., Macián, M.C., 2014. The family rhodobacteraceae. In: The Prokaryotes: Alphaproteobacteria and Betaproteobacteria. Springer Berlin Heidelberg, pp. 439–512.

Shchegolkova, N.M., Krasnov, G.S., Belova, A.A., Dmitriev, A.A., Kharitonov, S.L., Klimina, K.M., Melnikova, N.V., Kudryavtseva, A.V., 2016. Microbial community structure of activated sludge in treatment plants with different wastewater compositions. Front. Microbiol. 7, 90.

Stan Development Team, 2020. RStan: the R interface to Stan. http://mc-stan.org. r package version 2.19.31.

Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., Xavier, J.B., 2013. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput. Biol. 9, 1–11.

Volterra, V., 1926. Fluctuations in the abundance of a species considered mathematically. Nature 118, 558–560.

Wagner, M., Rath, G., Koops, H.P., Flood, J., Amann, R., 1996. In situ analysis of nitrifying bacteria in sewage treatment plants. Water Sci. Technol. 34, 237–244.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.

Wetzel, R.G., 2001. 12 - The nitrogen cycle. In: Wetzel, R.G. (Ed.), Limnology, 3 ed. Academic Press, San Diego, pp. 205–237.

Williams, K.P., Sobral, B.W., Dickerman, A.W., 2007. A robust species tree for the Alphaproteobacteria. J. Bacteriol., 4578–4586.

Xia, Y., Sun, J., Chen, D.G., 2018. Statistical Analysis of Microbiome Data with R. Springer Singapore, Singapore.

Xu, L., Xu, X., Kong, D., Gu, H., Kenney, T., 2020. Stochastic generalized Lotka-Volterra model with an application to learning microbial community structures. arXiv:2009.10922v1.

Zhang, J., Lin, W., 2019. Scalable estimation and regularization for the logistic normal multinomial model. Biometrics 75, 1098–1108.