Community detection in networks with unobserved edges

Till Hoffmann,[1] Leto Peel,[2] Renaud Lambiotte,[3] and Nick S. Jones[4]

[1] *Department of Mathematics, Imperial College, London SW7 2AZ, United Kingdom*
[2] *Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium*
[3] *Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom*
[4] *EPSRC Centre for Mathematics of Precision Healthcare, Imperial College, London SW7 2AZ, United Kingdom*

We develop a Bayesian hierarchical model to identify communities in networks for which we do not observe the edges directly, but instead observe a series of interdependent signals for each of the nodes. Fitting the model provides an end-to-end community detection algorithm that does not extract information as a sequence of point estimates but propagates uncertainties from the raw data to the community labels. Our approach naturally supports multiscale community detection as well as the selection of an optimal scale using model comparison. We study the properties of the algorithm using synthetic data and apply it to daily returns of constituents of the S&P100 index as well as climate data from US cities.

I. INTRODUCTION

Detecting communities in networks provides a means of coarse-graining the complex interactions or relations (represented by network edges) between entities (represented by nodes) and offers a more interpretable summary of a complex system. However, in many complex systems the exact relationship between entities is either unknown or unobserved. Instead, we may observe interdependent signals from the nodes, such as time series, which we may use to infer these relationships. Over the past decade, a multitude of algorithms have been developed to group multivariate time series into communities with applications in finance [1–4], neuroscience [5, 6], and climate research [7]. For example, identifying communities of assets whose prices vary coherently can help investors gain a deeper understanding of the foreign exchange market [1, 2] or manage their market risk by investing in assets belonging to different communities [8]. Classifying regions of the brain into distinct communities allows us to predict the onset of psychosis [6] and learn about the ageing of the brain [9]. Global factors affecting our climate are reflected in the community structure derived from sea surface temperatures [7].

Current methods for detecting communities when network edges are unobserved, typically involve a complicated process that is highly sensitive to specific design decisions and parameter choices. Most approaches consist of three steps: First, a measure is chosen to assess the similarity of any pair of time series such as Pearson correlation [1–3, 7, 9, 10], partial correlation [6, 11, 12], mutual information [13], or wavelet correlation coefficients [5, 14, 15]. Second, the similarity is converted to a dense weighted network [1–3, 15] or a binary network. For example, some authors connect the most similar time series such as to achieve a desired network density [13], threshold the similarity matrix at a single value [5, 7], or demand statistical significance under a null model [6, 11, 12]. Others threshold the similarity matrix at multiple values to perform a sensitivity analysis [9, 10, 14]. After the underlying network has been inferred, community detection is applied to uncover clusters of time series, for example by maximising the modularity [1, 2, 5, 10, 14, 15] or using the map equation [7, 9, 16].

This type of approach faces a number of challenges: first, most community detection methods rely on the assumption that the network edges have been accurately observed [17]. In addition, Newman-Girvan modularity [18], a popular measure to evaluate community structure in networks, is based on comparing the network to a null model that does not apply to networks extracted from time series data [8]. Second, when the number of time series is large, computing pairwise similarities is computationally expensive, and the entries of the similarity matrix are highly susceptible to noise. For example, the sample covariance matrix does not have full rank when the number of observations is smaller than or equal to the number of time series [19]. Third, at each step of the three-stage process we generally only compute point estimates and discard any notion of uncertainty such that it is difficult to distinguish genuine community structure from noise—a generic problem in network science [20]. Fourth, missing data can make it difficult to compute similarity measures such that data have to be imputed [10] or incomplete time series are dropped [3, 8]. Finally, and more broadly, determining an appropriate number of communities is difficult [21] and often relies on the tuning of resolution parameters without a quality measure to choose one value over another [2, 22].

Our approach is motivated by the observation that inferring the presence of edges between all pairs of nodes in a network is an unnecessary, computationally expensive step to uncover the presence of communities. Instead, we propose a Bayesian hierarchical model for multivariate time series data that provides an end-to-end community

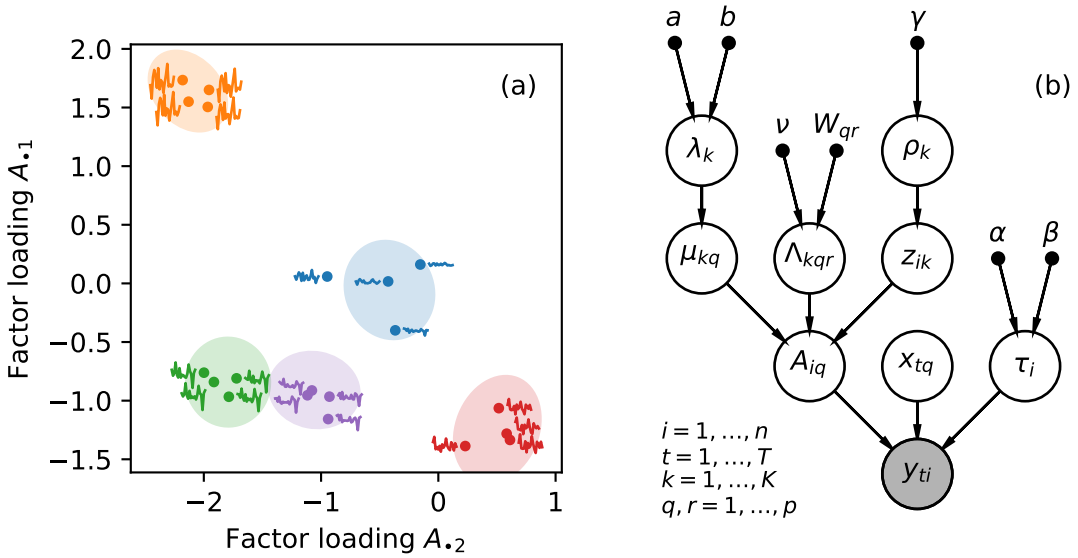

FIG. 1. A Bayesian hierarchical model for time series with community structure. Time series y are generated by a latent factor model with factor loadings A shown as dots in panel (a). The factor loadings are drawn from a Gaussian mixture model with mean μ and precision Λ. Generated time series are shown next to each factor loading for illustration. Panel (b) shows a directed acyclic graph representing the mixture model (A and all of its parents) and the probabilistic principal component analysis (A, its siblings, and y). Observed nodes are shaded grey and fixed hyperparameters are shown as black dots.

detection algorithm and propagates uncertainties directly from the raw data to the community labels. This shortcut is more than a computational trick, as it naturally allows us to address the aforementioned challenges. In particular, our approach naturally supports multiscale community detection as well as the selection of an optimal scale using model comparison. Furthermore, it enables us to extract communities even in the case of short observation time windows. The rest of this paper will be organised as follows. After introducing the algorithm, we validate and study its properties in a series of synthetic experiments. We then apply it to daily returns of constituents of the S&P100 index to identify salient communities of similar stocks and to climate data of US cities to identify homogeneous climate zones. For the latter, we characterise the quality of the communities in terms of the predictive performance provided by the model.

II. METHODS

The variability of high-dimensional time series is often the result of a small number of common, underlying factors [23]. For example, the stock price of oil and gas companies tends to be positively affected by rising oil prices, whereas the manufacturing industry, which consumes oil and gas, is likely to suffer from rising oil prices [24]. Motivated by this observation, we model the multivariate time series y using a latent factor model, i.e. the n-dimensional observations at each time step t are generated by a linear transformation A of a lower-dimensional, latent time series x and additive observation noise. More formally, the conditional distribution of y is

$$y_{ti}|A, x, \tau \sim \text{Normal}\left(\sum_{q=1}^{p} x_{tq} A_{iq}, \tau_i^{-1}\right) \tag{1}$$

where y_{ti} is the value of the i^{th} time series at time t, x_{tq} is the value of the q^{th} latent time series, and p is the number of latent time series. The precision (inverse variance) of the additive noise for each time series is τ_i, and $\text{Normal}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The entries A_{iq} of the $n \times p$ factor loading matrix encode how the observations of time series i are affected by the latent factor q. Using our earlier example, the entry of A connecting an oil company with the (unobserved) oil price would be positive, whereas the corresponding entry for an automobile company would be negative.

Variants of this model abound. For example, the mixture model of factor analysers [25, 26] assumes that there is not one but many latent factors to account for a possibly non-linear latent manifold [27, 28]. Huopaniemi *et al.* [29], Zhao *et al.* [30] demand that most of the entries of the factor loading matrix are zero such that each observation only depends on a subset of the latent factors. Inoue *et al.* [31] model gene expression data and assume that the factor loadings of all genes belonging to the same community are identical.

We aim to strike a balance between the restrictive assumption that observations belonging to the same community have identical factor loadings [31] and the more complex mixtures of factor analysers [27]: we define a community of time series as having factor loadings drawn from a common latent distribution. Each time series i belongs to exactly one community $g_i \in \{1, \ldots, K\}$, i.e. g is the vector of community memberships and K is the number of communities. The factor loadings are drawn from a multivariate normal distribution conditional on the community membership of each time series such that

$$A_i \sim \sum_{k=1}^{K} z_{ik} \text{Normal}\left(\mu_k, \Lambda_k^{-1}\right), \tag{2}$$
$$\text{where } z_{ik} = \begin{cases} 1 & \text{if } g_i = k \\ 0 & \text{otherwise} \end{cases}.$$

The parameters μ_k and Λ_k are the p-dimensional mean and precision matrix of the k^{th} component, respectively. The intuition behind the model is captured in panel (a) of fig. 1: we can identify communities because time series that behave similarly are close in the space spanned by the factor loading matrix. This idea relates to latent space models of networks in which nodes that are positioned closer together in the latent space have a higher probability of being linked [32]. Extending the notion of communities to such a model implies clusters of nodes within the latent space [33].

The priors for the mean and precision parameters of the different communities require careful consideration because they can have a significant impact on the outcome of the inference [34]: if the priors are too broad, the model evidence is penalised heavily for each additional community, and all time series are assigned to a single community. If the priors are too narrow, the inference will fail because it is dominated by our prior beliefs rather than being data driven. To minimise the sensitivity of our model to prior choices, we use an automatic relevance determination (ARD) prior, which can learn an appropriate scale for the centres of the communities μ [35]. In particular,

$$\mu_{kq} \sim \text{Normal}(0, \lambda_{kq}^{-1})$$
$$\lambda_{kq} \sim \text{Gamma}(a = 10^{-3}, b = 10^{-3}).$$

Conjugate ARD priors are not available for the precision matrices of the communities, and we use Wishart priors such that

$$\Lambda_k \sim \text{Wishart}(\nu, W I_p),$$

where $\nu > p - 1$ and W are the shape and scale parameters of the Wishart distribution, respectively, and I_p is the p-dimensional identity matrix. To obtain a relatively broad prior [36], we let $\nu = p$ such that the *prior precision*, i.e. the expectation of the precision under the prior, is $\langle \Lambda \rangle = W^{-1} I_p$. We will perform inference for a range of prior precisions because we cannot learn it automatically using an ARD prior.

Latent factor models as defined in eq. (1) are not uniquely identifiable because we can obtain an equivalent solution by, for example, multiplying the factor loading matrix A by an arbitrary constant and dividing the latent factors x by the same value. We impose a zero-mean, unit-variance Gaussian prior on the latent factors to identify the scale of x and A [37]. This approach does not identify the model with respect to rotations and reflections. But the lack of identifiability does not affect the detection of communities because the Gaussian mixture model defined in eq. (2) is invariant to orthogonal transformations.

The community memberships follow a categorical distribution

$$g_i \sim \text{Categorical}(\rho),$$

where ρ represents the normalised sizes of communities such that $\sum_{k=1}^{K} \rho_k = 1$. The community sizes have a Dirichlet prior

$$\rho \sim \text{Dirichlet}(\gamma),$$

where $\gamma = 10^{-3}$ is a uniform concentration parameter for all elements of the Dirichlet distribution such that no community is favoured a-priori. We use a broad Gamma prior for the precision parameter of the idiosyncratic noise. In particular,

$$\tau_i \sim \text{Gamma}(\alpha = 10^{-3}, \beta = 10^{-3}).$$

Panel (b) of fig. 1 shows a graphical representation of the model as a directed acyclic graph.

A. Inference using the variational mean-field approximation

Exact inference for the hierarchical model is intractable, and we use a variational mean-field approximation of the posterior distribution to learn the parameters [38]. The basic premise of variational inference is to approximate the posterior distribution $P(\Theta|y)$ by a simpler distribution $Q(\Theta)$, where Θ is the set of all parameters of the model. Variational inference algorithms seek the approximation $Q^*(\Theta)$ that minimises the Kullback-Leibler divergence between the approximation and the true posterior. More formally,

$$Q^*(\Theta) = \underset{Q \in \mathcal{Q}}{\text{argmin}}\, \text{KL}\left(Q(\Theta) \| P(\Theta|y)\right),$$

where $\mathcal{Q}$ is the space of all approximations we are willing to consider. Minimising the Kullback-Leibler divergence is equivalent to maximising the evidence lower bound (ELBO)

$$L(Q) = \langle \log P(y, \Theta) \rangle \leq \log \int d\Theta\, P(y, \Theta), \tag{3}$$

where $\langle \cdot \rangle$ denotes the expectation with respect to the approximate posterior Q and the right-hand side of eq. (3) is the logarithm of the model evidence [38]. The maximised ELBO (henceforth just ELBO) serves as a proxy for the model evidence to perform model comparison, and we will use it to determine the number of latent factors and the prior precision.

We further assume that the posterior approximation factorises with respect to the nodes of the graphical model shown in fig. 1 (a). More formally, we let $Q(\theta) = \prod_{\theta_i \in \Theta} Q_{\theta_i}(\theta_i)$ which restricts the function space $\mathcal{Q}$. Under this assumption, known as the mean-field approximation, the individual factors can be optimised in turn until the ELBO converges to a (local) maximum. The general update equation is (up to an additive normalisation constant)

$$\log Q_{\theta_i}(\theta_i) \rightarrow \langle \log P(\Theta|y) \rangle_{\setminus \theta_i},$$

where $\langle \cdot \rangle_{\setminus \theta_i}$ denotes the expectation with respect to all parameters except the parameter θ_i under consideration. See Blei *et al.* [39] for a recent review of variational Bayesian inference and appendix B for the update equations specific to our model.

III. SIMULATION STUDY

Having developed an inference algorithm for the model, we would like to assess under which conditions the algorithm fails and succeeds. We start with a simple, illustrative example by drawing $K = 5$ community means μ from a two-dimensional normal distribution with zero mean and unit variance, i.e. we consider two latent time series and a two-dimensional space of factor loadings. The community precisions Λ are drawn from a Wishart distribution with shape parameter $\nu = 50$ and identity scale parameter. The communities are well-separated because the within-community variability ($1/\sqrt{50} \approx 0.14$) is much smaller than the between-community variability (≈ 1) as shown in panel (a) of fig. 2. We assign $n = 50$ time series to the five communities using a uniform distribution of community sizes $\rho_k = 1/K$. Finally, we draw $m = 100$ samples of the two-dimensional latent factors x and obtain the observations y using eq. (1), i.e. by adding Gaussian observation noise with precision τ drawn from a Gamma$(100, 10)$ distribution to the linear transformation xA^T.

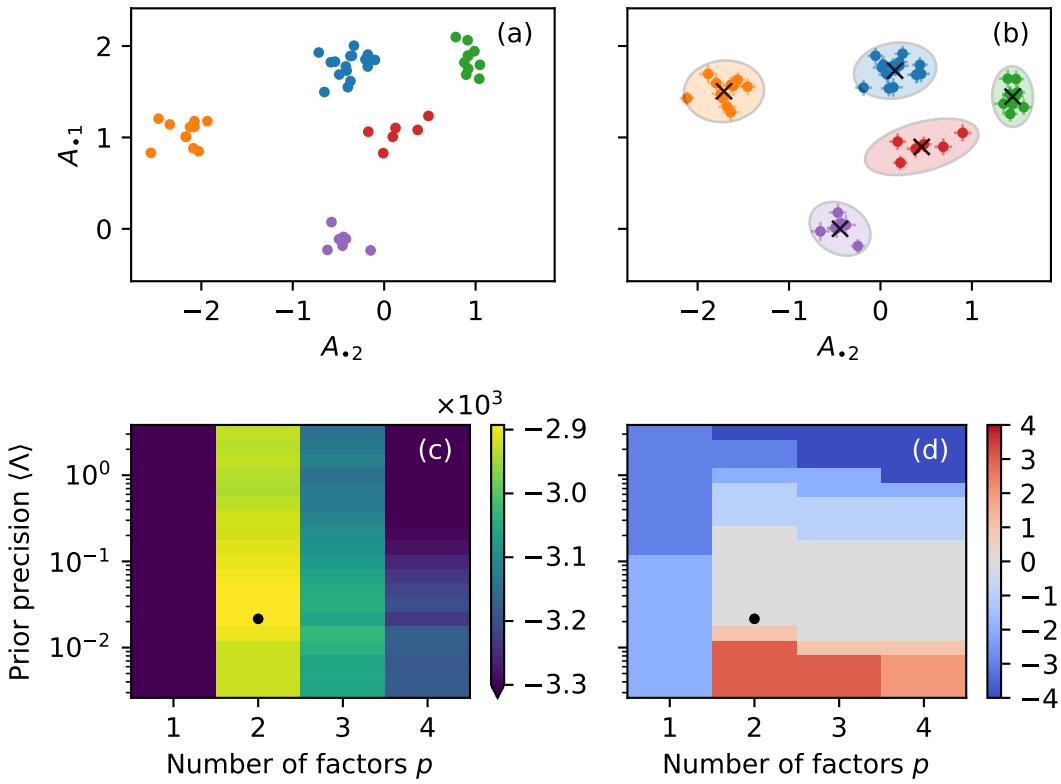

FIG. 2. The algorithm successfully identifies synthetic communities of time series. Panel (a) shows the entries of a synthetic factor loading matrix A as a scatter plot. Panel (b) shows the inferred factor loading matrix together with the community centres as black crosses and the community covariances as ellipses; error bars correspond to three standard deviations of the posterior. Panel (c) shows the ELBO as a function of the number of latent factors and the prior precision. Panel (d) shows the difference between the estimated number of communities $\hat{K}$ and the true number of communities K. The model with the highest ELBO is marked with a black dot in panels (c) and (d); it recovers two latent factors and five communities.

Optimising the ELBO is usually a non-convex problem [39], and the results are sensitive to the initalisation of the posterior factors. Choosing a good initialisation is difficult in general, but the optimisation can be aided to converge more quickly by initialising it using a simpler algorithm [40]. We run the inference algorithm in three stages: first, we fit a standard probablistic PCA [41] to initialise the latent factors, factor loadings, and noise precision. Second, we perform ten independent runs of k-means clustering on the factor loading matrix [42] and update the community assignments z according to the result of the best run of the clustering algorithm, i.e. the clustering with the smallest sum of squared distances between the factor loadings A and the corresponding cluster centres μ. Third, we optimise the posterior factors of all parameters according to the variational update equations in appendix B until the ELBO does not increase by more than a factor of 10^{-6} in successive steps. The entire process is repeated 50 times and we choose the model with the highest ELBO to mitigate the optimisation algorithm getting stuck in local optima.

The number of communities and the prior precision are tightly coupled: suppose we choose a large prior precision for the Wishart distribution encoding a prior belief that each individual community occupies a small volume in the space of factor loadings. Consequently, the algorithm is incentivised to separate the time series into many small communities. In the limit $\langle\Lambda\rangle \to \infty$ (where vanishing within-community variation is permitted), the algorithm assigns each time series to its own community. In contrast, if we choose a small prior precision, our initial belief is that each community occupies a large volume in the latent space, and time series are aggregated into few, large communities. Fortunately, the number of communities is determined automatically once the prior precision has been specified: in practice, we define the inferred cluster labels as

$$\hat{g}_i = \operatorname*{argmax}_k \langle z_{ik} \rangle_{Q_z} ,$$

and determine the number of inferred communities $\hat{K}$ by counting the number of unique elements in $\hat{g}$.

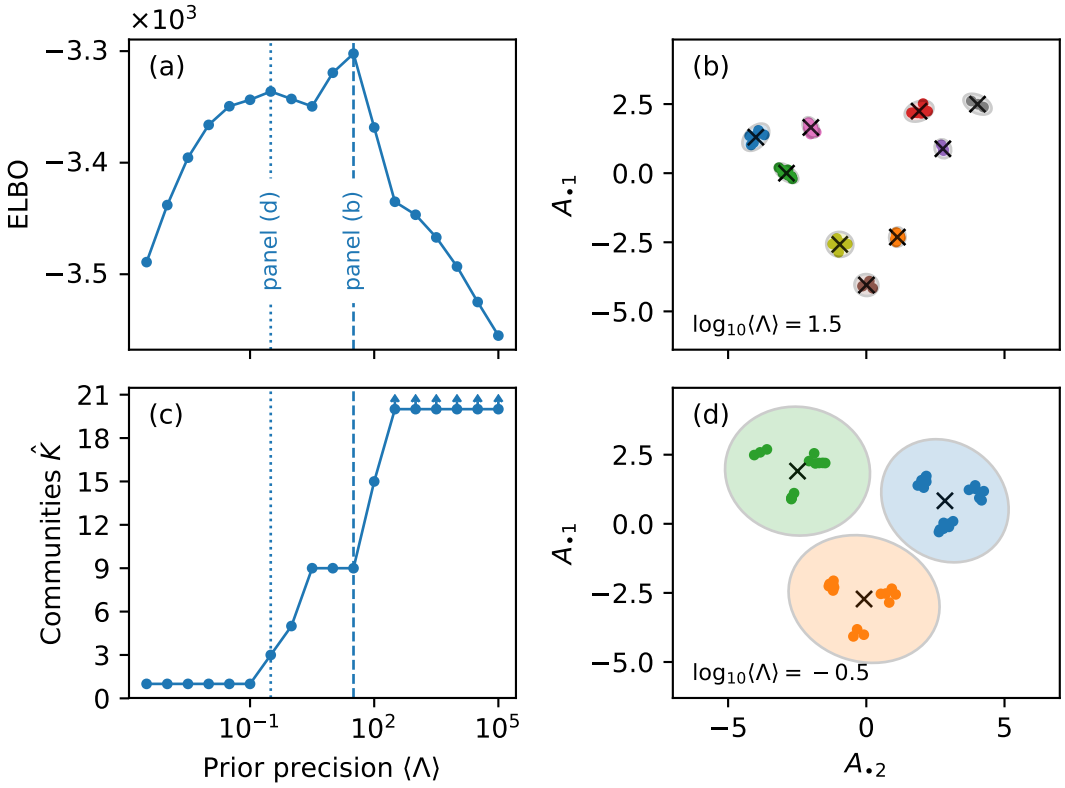

FIG. 3. The prior precision $\langle\Lambda\rangle$ of the communities affects the number of detected communities. Panel (a) shows the ELBO of the model as a function of the prior expectation of the precision matrices $\langle\Lambda\rangle$. The ELBO has two distinct peaks corresponding to the community assignments shown in panels (b) and (d), respectively. Panel (c) shows the number of identified communities as a function of the prior precision; data points with arrows represent a lower bound on the number of inferred communities.

For the synthetic data discussed above, we set the maximum number of communities to ten and run the inference for a varying number of latent factors and prior precisions. Increasing the maximum number of communities would not have any effect because the algorithm identifies at most eight communities. The ELBO of the best model for each parameter pair is shown in panel (c) of fig. 2. The model with the highest ELBO correctly identifies the number of factors and the number of communities; the inferred parameters are shown in panel (b). As mentioned in section II, the model is not identifiable with respect to rotations and reflections and consequently the factor loadings in panels (a) and (b) differ. However, the precise values do not affect the community assignments, and the difference is immaterial. Panel (d) shows the difference between the inferred and actual number of communities. As expected, choosing too small or large a prior precision leads to the algorithm inferring too few or many communities, respectively.

Choosing the hyperparameters, such as the number of factors and the prior precision, to maximise the ELBO is known as empirical Bayes [38]. In theory, it is preferable to introduce hyperpriors and treat the number of factors and the prior precision as proper model parameters similar to the ARD prior. However, dealing with the variable dimensionality of the latent space is difficult in practice and computationally convenient conjugate priors for the scale parameter of Wishart distributions do not exist.

A. Multiscale community detection

Treating the dimensionality p of the latent space and the extent Λ of communities in the latent space as input parameters not only lets us avoid complicated inference but also provides us with a natural approach to multiscale community detection. We create nine communities arranged in a hierarchical fashion in the factor loading space similar to a truncated Sierpiński triangle and assign $n = 50$ time series to the communities as shown in panel (b) of fig. 3. As in the previous section, we generate $T = 100$ observations of the time series with noise precision drawn from a Gamma$(100, 10)$ distribution.

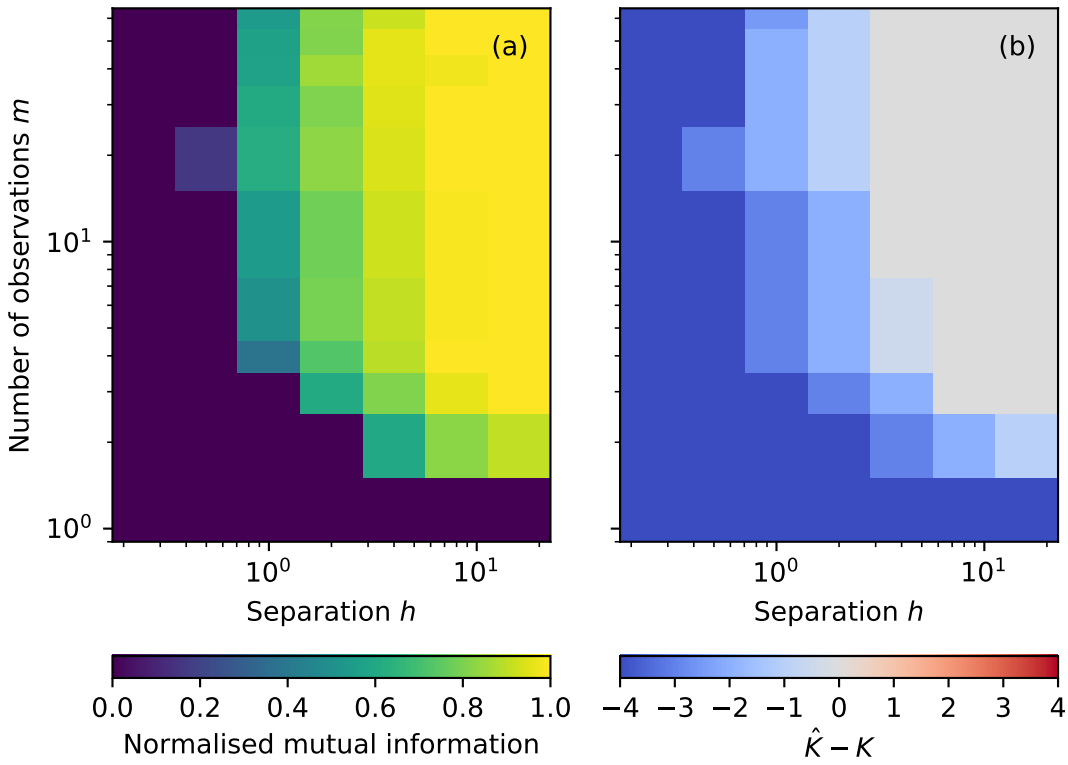

FIG. 4. Communities can be recovered even from very short time series. Panel (a) shows the median normalised mutual information between the true and inferred community assignments for $n = 50$ time series and $K = 5$ groups as a function of the number of observations T and the community separation h. Panel (b) shows the median difference between the number of inferred communities and the true number of communities.

In this example, we assume that the number of latent factors is known, set the maximum number of communities to 20, and vary the prior precision over several orders of magnitude. Panel (a) of fig. 3 shows the ELBO as a function of the prior precision exhibiting two local maxima: the larger of the two corresponds to a large prior precision and identifies the nine communities used to generate the data as shown in panel (b). The smaller maximum occurs at a smaller prior precision and the algorithm aggregates time series into mesoscopic communities as shown in panel (d). Decreasing the prior precision further forces the algorithm to assign all time series to a single community, and increasing the prior precision beyond its optimal value results in communities being fragmented into smaller components as can be seen in panel (c). Our algorithm is able to select an appropriate scale automatically but also allows the user to select a particular scale of interest if desired.

B. Testing the limits

In both of the examples we have considered so far, the communities were well separated from one another which made it easier to assign time series to communities. Similarly, the number of observations T was twice as large as the number of time series n such that the algorithm could constrain the factor loading matrix well. In this section, we consider how the performance of the algorithm changes as we change the separation between communities and the number of observations. We define the community separation

$$h = \sqrt{\langle \Lambda \rangle \, \mathrm{var}[\mu]}$$

which measures the relative between-community and within-community scales such that communities are well-separated in the factor loading space if $h \gg 1$, and are overlapping if $h \ll 1$. The expectation and variance in the definition of h are taken with respect to the generative model for the synthetic data.

For each combination of the number of observations and the separation h, we run 50 independent simulations with $K = 5$ communities, prior precision $\Lambda = 10I_2$ for each community, and $p = 2$ latent factors. For the inference, we assume that the number of latent factors is known and impose a limit of at most 10 communities. The prior precision is varied logarithmically from 0.625 to 20, and we retain the model with the highest ELBO. We use two criteria to measure the performance of the algorithm.

First, we measure the normalised mutual information (NMI) between the inferred community labels $\hat{g}$ and the true community labels g. The NMI is equal to one if the inferred and true community labels match exactly and is equal to zero if the community labels are independent. It is defined as [43]

$$\mathrm{NMI}(g, \hat{g}) = \frac{I(g, \hat{g})}{\sqrt{H(g)H(\hat{g})}},$$

where $I(g, \hat{g})$ is the mutual information between the true and inferred community assignments, and $H(g)$ is the entropy of g. The NMI displayed in panel (a) of fig. 4 shows a clear and expected pattern: the larger the separation and the larger the number of observations, the better the inference. The separation poses a fundamental limit to how well we can infer the community labels. Even if we could estimate the factor loadings perfectly, we could not determine the community memberships if the communities are overlapping. This observation is analogous to the detectability limit for community detection on fully-observed networks: the ability to recover community assignments diminishes as the difference of within-community and between-community connections decreases [44]. However, provided that the communities are well separated, we can estimate the community labels well with a relatively small number of observations. We only require that the estimation error of the factor loadings are small compared to the separation between communities. Of course, the community separation is not under our control in practice, so we should ensure that we collect enough data to estimate the factor loadings well.

Second, we compare the inferred number of communities $\hat{K}$ with the true number of planted communities as shown in panel (b) of fig. 4. When the communities are overlapping, the algorithm infers a smaller number of communities because aggregating time series into fewer communities with more constituents provides a more parsimonious explanation of the data. Similarly, when the number of observations is too small, the factor loadings are not estimated well, and the algorithm chooses fewer communities because the data do not provide sufficient evidence to split the set of time series into smaller communities.

IV. APPLICATION TO FINANCIAL TIME SERIES

Having studied the behaviour of the algorithm on synthetic data, we apply it to daily returns of constituents of the S&P100 index comprising 102 stocks of 100 large companies in the United States. Google and 21$^{\mathrm{st}}$ Century Fox have two classes of shares and we discard FOXA and GOOG in favour of FOX and GOOGL, respectively, because the latter have voting rights. We obtained 252 daily closing prices for all stocks from 4$^{\mathrm{th}}$ of January to 30$^{\mathrm{th}}$ of December 2016 from Yahoo! finance[1]. Before feeding the data to our algorithm, we compute the daily logarithmic returns for each time series and standardise them by subtracting the mean and dividing by the standard deviation.

In contrast to performing a grid search over the number of latent factors and the prior precision jointly as in section III, we run the inference in two steps. First, we fit a standard probabilistic PCA model [41] and use the ELBO to choose the number of latent factors as shown in panel (a) of fig. 5. Having identified the optimal number of factors as $\hat{p} = 10$, we perform a grid search over the prior precision to select an appropriate scale for the communities. The algorithm selects $\hat{K} = 11$ communities as shown in panels (b) and (c) of fig. 5. Amongst an ensemble of 50 independently fitted models for each prior precision, the model with the highest ELBO tends to have the smallest number of communities: the algorithm tries to find a parsimonious description of the data and representations with too many communities are penalised.

The factor loading matrix A has non-trivial structure as can be seen in panel (a) of fig. 6: the columns of the factor loading matrix are ordered descendingly according to the column-wise L_2 norm. The first column explains most of the variance of the data and the corresponding factor is often referred to as the market mode which captures the overall sentiment of investors [8, 45]. Additional factors capture ever more refined structure. Because visualising the ten-dimensional factor loading matrix is difficult, we obtain a lower-dimensional embedding using t-SNE [46] shown in panel (b). The shaded regions are the convex hulls of time series belonging to the same community.

The community assignments capture salient structure in the data. For example, the three smallest communities that have only two members consist of: MasterCard (MA) and Visa (V), both credit card companies; Lockheed

[1] `https://finance.yahoo.com/`

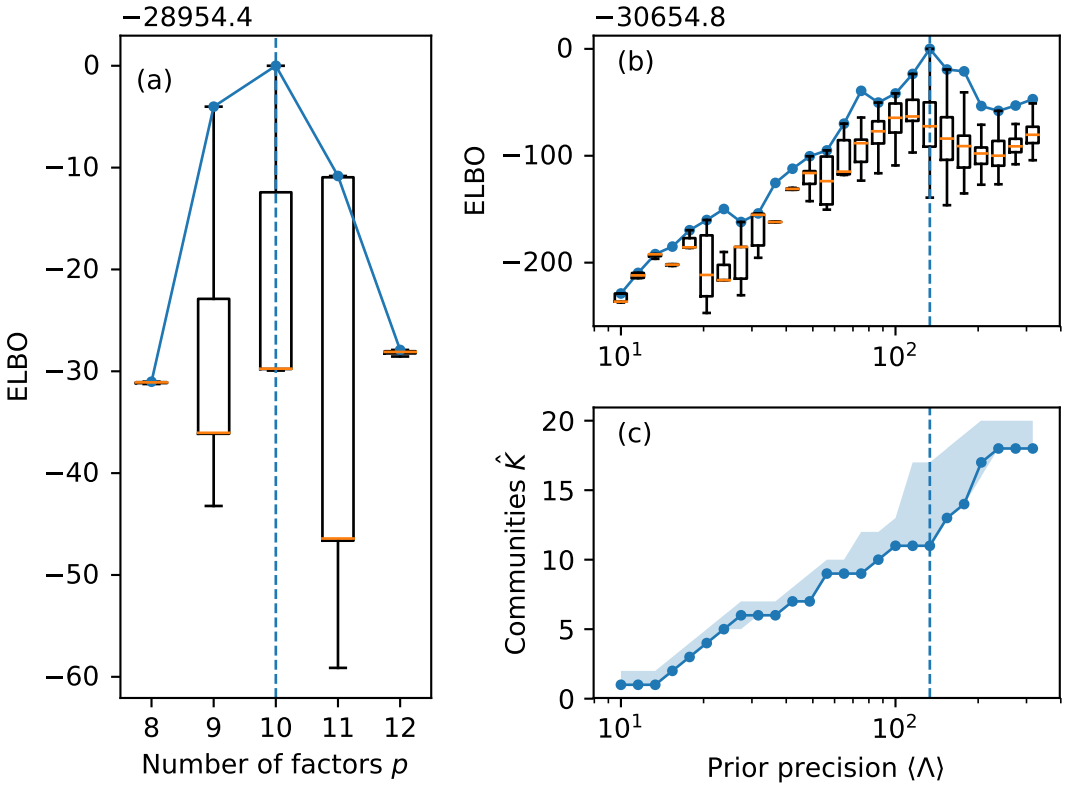

FIG. 5. The algorithm identifies eleven communities of stocks in a ten-dimensional factor loading space. Panel (a) shows the ELBO as a function of the number of latent factors of the model peaking at $p = 10$ factors. The ELBO of the best of an ensemble of 50 independently fitted models is shown in blue; the box plot provides a measure of the variability induced by starting the algorithm from different initial conditions. Panel (b) shows the ELBO as a function of the prior precision, and panel (c) shows the number of communities identified by the algorithm. The shaded region corresponds to the range of the number of detected communities in the model ensemble.

Martin (LMT) and Raytheon (RTN), both defence companies; and DuPont (DD) and Dow Chemical (DOW), both chemical companies. Dow Chemical and DuPont merged to form the conglomerate DowDuPont (DWDP) in August 2017. The algorithm also identifies a large community of companies from diverse industry sectors. More specialised communities consist of biotechnology and pharmaceutical companies (e.g. Merck (MRK), Gilead Sciences (GILD)), financial services companies (e.g. Citigroup (C), Goldman Sachs (GS)), as well as manufacturing and shipping (e.g. Boeing (BA), Caterpillar (CAT), FedEx (FDX), United Parcel Service (UPS)).

Some of the community assignments appear to be less intuitive. For instance, the nuclear energy company Exelon (EXC) is assigned to a community of telecommunications companies rather than to a community of other energy companies as we might expect. This result does not necessarily indicate an error in community assignment, as the "true" communities in real data are not known [47]. However, in this case we see in the t-SNE embedding that Exelon is closer to other energy companies, which suggests that this particular assignment may be an artefact of assuming Gaussian distributed factor loadings. See table I for a full list of companies and community assignments.

V. APPLICATION TO CLIMATE DATA

We now apply our method to climate data from 1,429 US cities[2]. Each "node" represents a city, and the signals we observe at each of the nodes are monthly values (averaged over 20 years) for the high and low temperatures and the amount of precipitation received. So instead of T observations of a time series, we have T attributes of the nodes,

[2] Data downloaded from `https://www.usclimatedata.com`.

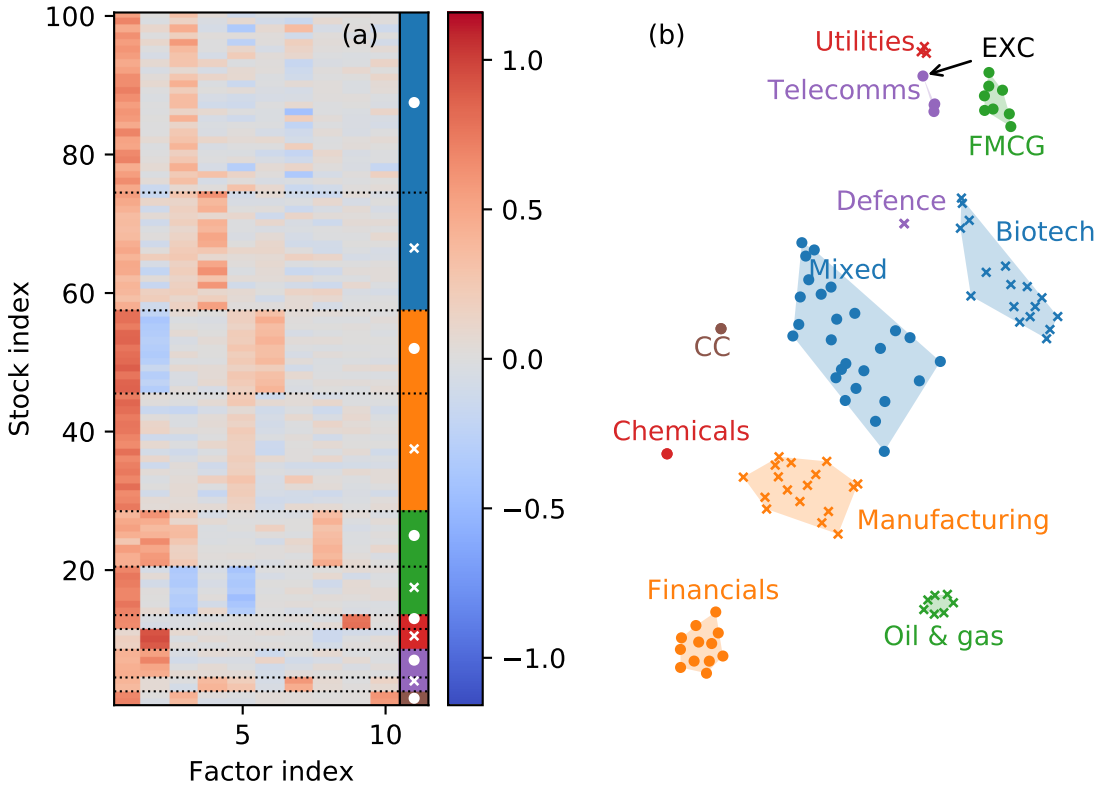

FIG. 6. The detected communities of stocks are correlated with industry sectors. Panel (a) shows the factor loadings inferred from one year of daily log-returns of constituents of the S&P100 index as a heat map. Each row corresponds to a stock and each column corresponds to a factor. The last column of the loading matrix serves as a colour key for different communities. Panel (b) shows a two-dimensional embedding of the factor loading matrix using t-SNE together with cluster labels including credit card (CC) and fast-moving consumer goods (FMCG) companies. See the main text for a discussion of the nuclear energy company Exelon (EXC).

in this case $T = 36$ (three times twelve months). In this context, communities represent climate zones in which the temperature and precipitation vary similarly. In climatology, locales are classified into climate zones according to man-made climate classification schemes. One of the most popular climate classification schemes is the Köppen-Geiger climate classification system [48], first developed in 1884 by Wladimir Köppen [49], but has since received a number of modifications. The system divides climates into groups based on seasonal temperature and precipitation patterns. Figure 7 (a) shows the Köppen-Geiger classification of the US cities we studied.

We infer the parameters of our model and community assignments using a similar approach to the previous section except for two notable differences. First, we found that the ELBO increased monotonically with increasing number of latent factors when fitting the standard probabilistic PCA. We thus decided to use six latent factors as the rate of increase of the ELBO drops when we increase the number of factors further. Second, instead of choosing the number of communities by maximising the ELBO, we set the number of communities to the number of Köppen-Geiger climate zones to allow for a more direct comparison. Figure 7 (b) shows the communities inferred by our model. Both sets of climate zones display similar qualitative features such as the division between the humid East and the arid West along the 100^{th} meridian. However, a direct quantitative comparison of the two climate partitions is not necessarily meaningful as we do not expect there to be only a single good way to partition the nodes. For reference, we find the normalised mutual information between the two community assignments is ≈ 0.4. The low correlation between our inferred communities and the manually labelled Köppen-Geiger zones does not imply poor performance of our model [47], but nor does it validate it.

Instead of trying to recover man-made labels, we consider the predictive performance of our model on held-out, previously unseen data. We first fit the model to the complete series of 760 cities selected uniformly at random, which acts as a training set to learn the latent factors x, community means μ, and community precisions Λ. Second, we perform a ten-fold cross validation on the remaining 669 cities by holding out a tenth of the data, inferring their factor

Group	Constituents
Mixed	Apple (AAPL), Abbott Laboratories (ABT), Accenture (ACN), Amazon (AMZN), American Express (AXP), Cisco (CSCO), Danaher (DHR), Walt Disney (DIS), Facebook (FB), Twenty-First Century Fox (FOX), Google (GOOGL), Home Depot (HD), Intel (INTC), Lowe's (LOW), Medtronic (MDT), Monsanto (MON), Microsoft (MSFT), Nike (NKE), Oracle (ORCL), Priceline.com (PCLN), Paypal (PYPL), Qualcomm (QCOM), Starbucks (SBUX), Time Warner (TWX), Texas Instruments (TXN), Walgreen (WBA)
Biotech	AbbVie (ABBV), Actavis (AGN), Amgen (AMGN), Biogen (BIIB), Bristol-Myers Squibb (BMY), Celgene (CELG), Costco (COST), CVS (CVS), Gilead (GILD), Johnson & Johnson (JNJ), Eli Lilly (LLY), McDonald's (MCD), Merck (MRK), Pfizer (PFE), Target (TGT), UnitedHealth (UNH), Walmart (WMT)
Financials	American International Group (AIG), Bank of America (BAC), BNY Mellon (BK), BlackRock (BLK), Citigroup (C), Capital One (COF), Goldman Sachs (GS), JPMorgan Chase (JPM), MetLife (MET), Morgan Stanley (MS), US Bancorp (USB), Wells Fargo (WFC)
Manufacturing & shipping	Allstate (ALL), Barnes Group (B), Boeing (BA), Caterpillar (CAT), Comcast (CMCSA), Emerson Electric (EMR), Ford (F), FedEx (FDX), General Dynamics (GD), General Electric (GE), General Motors (GM), Honeywell (HON), International Business Machines (IBM), 3M (MMM), Union Pacific (UNP), United Parcel Service (UPS), United Gechnologies (UTX)
Fast-moving consumer goods	Colgate-Palmolive (CL), Kraft Heinz (KHC), Coca Cola (KO), Mondelez International (MDLZ), Altria (MO), PepsiCo (PEP), Procter & Gamble (PG), Philip Morris International (PM)
Oil & gas	ConocoPhillips (COP), Chevron (CVX), Halliburton (HAL), Kinder Morgan (KMI), Occidental Petroleum (OXY), Schlumberger (SLB), ExxonMobil (XOM)
Chemicals	DuPont (DD), Dow Chemical (DOW)
Utilities	Duke Energy (DUK), Nextera (NEE), Southern Company (SO)
Telecomms	Exelon (EXC), Simon Property Group (SPG), AT&T (T), Verizon (VZ)
Defence	Lockheed Martin (LMT), Raytheon (RTN)
Credit cards	MasterCard (MA), Visa (V)

TABLE I. Constituents of the S&P100 grouped by inferred community assignment.

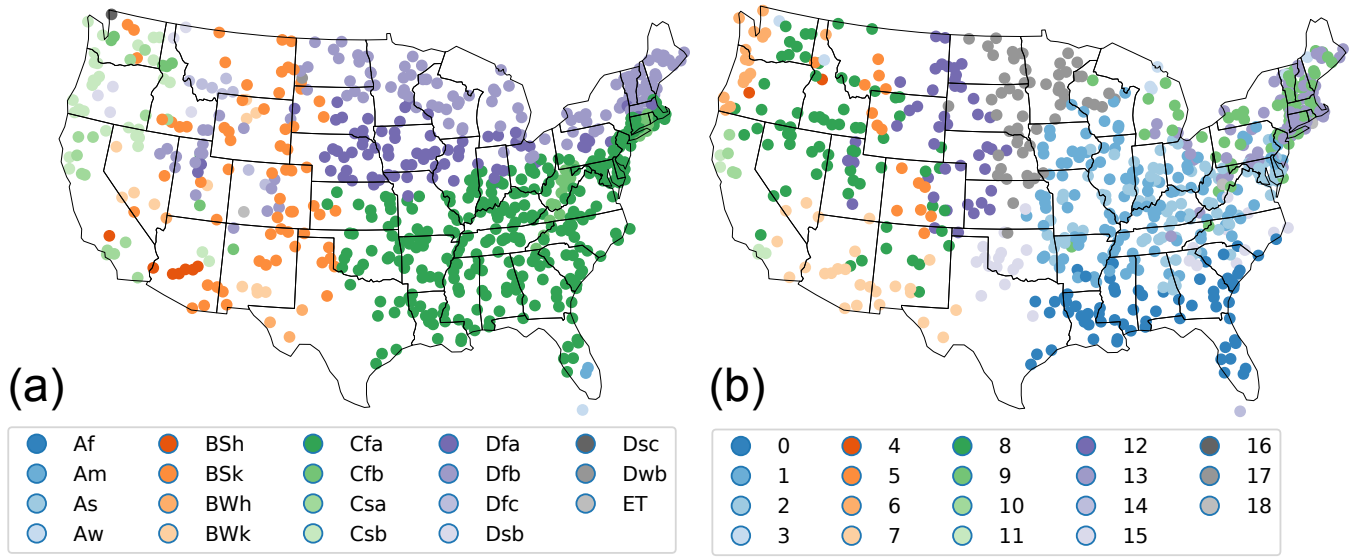

FIG. 7. Climate zones of US cities. Panel (a) shows the city locations coloured according to the Köppen-Geiger climate classification system [48]. Panel (b) shows the inferred climate zones based on the monthly average high and low temperatures and precipitation amounts. We observe qualitative similarities between the two sets of climate zones, but a quantitative comparison reveals a relatively low correlation (NMI ≈ 0.4).

loadings A and their community assignment, and predicting the missing signal values as $\hat{y}_i = A_i^T x$.

For comparison, we impute the missing values using the mean value of each signal type, i.e. the mean temperature or precipitation for each month, within each Köppen-Geiger climate zone [48]. Similarly, we make predictions using the communities found by a typical network-based method for clustering time series [2]. In particular, we apply the Louvain algorithm [50] with resolution parameter $\gamma = 0.95$ (so that we get approximately the same number of

Our method $A_i^T x$	**Our method** $\mu_{g_i}^T x$	**Köppen-Geiger [48]**	**Fenn *et al.* [2]**
0.301	0.578	0.706	0.727

TABLE II. RMSE predicting held-out climate signals.

communities as Köppen-Geiger zones) to the weighted adjacency matrix

$$A_{ij} = \frac{\rho_{ij} + 1}{2} - \delta_{ij},$$

where ρ_{ij} is the Pearson correlation between series i and j, and the Kronecker delta δ_{ij} removes self edges [2]. To provide a more direct comparison with the communities found by our method, we also compare the predictions using the community means, i.e. $\hat{y}_i = \mu_{g_i}^T x$.

Table II shows the root mean squared error (RMSE) for each approach. Our method outperforms the other two in terms of predictive ability. Whilst this observation provides some validation of our approach, it should not come as a surprise that our *data-driven* method, which is trained on the same type of data we are trying to predict, outperforms the hand-crafted zones of Köppen and Geiger. However, the approach detecting communities using a weighted adjacency matrix [2] performs worse than the Köppen-Geiger climate zones despite being trained on the same data: the method may identify spurious communities—at least with respect to those that have good predictive performance.

VI. DISCUSSION

We have developed a model for community detection for networks in which the edges are not observed directly. Using a series of interdependent signals observed for each of the nodes, our model detects communities using a combination of a latent factor model, which provides a lower-dimensional latent-space embedding, and a Gaussian mixture model, which captures the community structure. We fit the model using a Bayesian variational mean-field approximation which allows us to determine the number of latent factors as well as an appropriate number of communities using the ELBO for model comparison. The method is able to recover meaningful communities from daily returns of constituents of the S&P100 index and climate data in US cities. The code to run the inference is publicly available[3].

Our proposed method presents an important advancement over current methods for detecting communities when network edges are unobserved. Recall that these methods typically consist of three steps: calculate pairwise similarity, threshold similarity to create a network, and apply community detection to the network. In contrast, our approach is end-to-end, i.e. the method propagates uncertainties from the raw data to the community labels instead of relying on a sequence of point estimates. As a result, the model is able to recover community structure even when the number of observations T is possibly much smaller than the number of n. Current methods for detecting communities when network edges are unobserved struggle in this setting because of the uncertainty in the estimate of the similarity matrix. The asymptotic complexity of algorithms that rely on pairwise similarities scales (at least) quadratically with the number of nodes whereas each iteration of our algorithm scales linearly.

There are several avenues for future work. For example, using the same prior precision for all communities reflects our prior belief that all communities should occupy roughly similar volumes in the factor loading space. In analogy, in the case of standard community detection with modularity optimisation, balanced sizes between communities are induced by the so-called diversity index in the quality function [51]. Whether this assumption holds in practice is unclear, and we may be able to discover communities of heterogeneous sizes in the factor loading space by lifting this assumption. Furthermore, Gaussian distributions are a standard choice for mixture models, but mixtures of other distributions such as student-t distributions may provide better clustering results.

Despite being motivated by time series, our algorithm does not model the dynamics of the data explicitly. Using a dynamical model such as a linear state space model may capture additional information in the data to help infer better community labels and allow us to predict future values of the time series.

As shown in section V, our algorithm can recover communities from observations of different attributes. Whilst this use of the model violates the assumption that node observations are identically distributed, it does not prevent us from identifying meaningful communities. However, it may perform poorly in a posterior predictive check that compares statistics of the posterior distribution $P(y'|y)$ with the observed data. Promoting the observations y and

[3] `https://github.com/tillahoffmann/time_series/`

factor loadings A to three-dimensional tensors would allow us to model different attributes in a principled fashion. In particular, the l^{th} attribute of node i at time t would have distribution

$$y_{til}|A, x, \tau \sim \text{Normal}\left(\sum_{q=1}^{p} x_{tq}A_{ilq}, \tau_{il}^{-1}\right),$$

where A_{ilq} controls the effect of the q^{th} latent factor on attribute l of node i. Whilst increasing the number of independent observations T can only help us constrain the factor loadings A, collecting data about additional attributes provides fundamentally new information. Provided that the community assignments for the Gaussian mixture model are shared across the factor loadings of different attributes, we would be able to assign nodes to the correct community even if the components are not resolvable independently, i.e. $h \ll 1$, as discussed in section III B—similar to the enhanced detectability of fixed communities in temporal [52] and multilayer [53] networks.

Finally, this work provides a new perspective on how to perform network-based measurements in empirical systems where edges are not observed. This opens the way to other end-to-end methods for, e.g. estimating centrality measures or motifs in complex dynamical systems.

[1] Daniel J. Fenn, Mason A. Porter, Mark McDonald, Stacy Williams, Neil F. Johnson, and Nick S. Jones, "Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis," Chaos **19**, 033119 (2009).

[2] Daniel J. Fenn, Mason A. Porter, Peter J. Mucha, Mark McDonald, Stacy Williams, Neil F. Johnson, and Nick S. Jones, "Dynamical clustering of exchange rates," Quantitative Finance **12**, 1493–1520 (2012).

[3] Marya Bazzi, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison, "Community detection in temporal multilayer networks, with an application to correlation networks," Multiscale Modeling & Simulation **14**, 1–41 (2016).

[4] Tomohiro Ando and Jushan Bai, "Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures," Journal of the American Statistical Association **0**, 1–17 (2017).

[5] David Meunier, Renaud Lambiotte, Alex Fornito, Karen D Ersche, and Edward T Bullmore, "Hierarchical modularity in human brain functional networks," Frontiers in Neuroinformatics **3**, 37 (2009).

[6] Louis-David Lord, Paul Allen, Paul Expert, Oliver Howes, Matthew Broome, Renaud Lambiotte, Paolo Fusar-Poli, Isabel Valli, Philip McGuire, and Federico E. Turkheimer, "Functional brain networks before the onset of psychosis: A prospective fmri study with graph theoretical analysis," NeuroImage: Clinical **1**, 91–98 (2012).

[7] A. Tantet and H. A. Dijkstra, "An interaction network perspective on the relation between patterns of sea surface temperature variability and global mean surface temperature," Earth System Dynamics **5**, 1–14 (2014).

[8] Mel MacMahon and Diego Garlaschelli, "Community detection for correlation matrices," Phys. Rev. X **5**, 021006 (2015).

[9] Micaela Y. Chan, Denise C. Park, Neil K. Savalia, Steven E. Petersen, and Gagan S. Wig, "Decreased segregation of brain systems across the healthy adult lifespan," Proceedings of the National Academy of Sciences **111**, 4997–5006 (2014).

[10] Sen Wu, Mengjiao Tuo, and Deying Xiong, "Community structure detection of shanghai stock market based on complex networks," in *4th International Conference on Logistics, Informatics and Service Science* (2015) pp. 1661–1666.

[11] Anand S Pandit, Paul Expert, Renaud Lambiotte, Valerie Bonnelle, Robert Leech, Federico E Turkheimer, and David J Sharp, "Traumatic brain injury impairs small-world topology," Neurology **80**, 1826–1833 (2013).

[12] Tianwei Yu and Yun Bai, "Network-based modular latent structure analysis," BMC Bioinformatics **15**, S6 (2014).

[13] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "The backbone of the climate network," Europhysics Letters **87**, 48007 (2009).

[14] Aaron Alexander-Bloch, Renaud Lambiotte, Ben Roberts, Jay Giedd, Nitin Gogtay, and Ed Bullmore, "The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia," NeuroImage **59**, 3889–3900 (2012).

[15] R. F. Betzel, T. D. Satterthwaite, J. I. Gold, and D. S. Bassett, "A positive mood, a flexible brain," arXiv , 1601.07881 (2016).

[16] Martin Rosvall and Carl T Bergstrom, "Maps of random walks on complex networks reveal community structure," Proceedings of the National Academy of Sciences **105**, 1118–1123 (2008).

[17] Santo Fortunato, "Community detection in graphs," Physics Reports **486**, 75–174 (2010).

[18] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E **69**, 026113 (2004).

[19] Tony Cai, Weidong Liu, and Xi Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," Journal of the American Statistical Association **106**, 594–607 (2011).

[20] M. E. J. Newman, "Measurement errors in network data," arXiv , 1703.07376 (2017).

[21] P Latouche, E Birmelé, and C Ambroise, "Variational bayesian inference and complexity control for stochastic block models," Statistical Modelling **12**, 93–115 (2012).

[22] Jörg Reichardt and Stefan Bornholdt, "Statistical mechanics of community detection," Phys. Rev. E **74**, 016110 (2006).

[23] Eugene F. Fama and Kenneth R. French, "Common risk factors in the returns on stocks and bonds," Journal of Financial Economics **33**, 3–56 (1993).

[24] Mohan Nandha and Robert Faff, "Does oil move equity prices? a global view," Energy Economics **30**, 986–997 (2008).

[25] Zoubin Ghahramani and Geoffrey E. Hinton, *The EM Algorithm for Mixtures of Factor Analyzers*, Tech. Rep. CRG-TR-96-1 (University of Toronto, 1996).

[26] Zoubin Ghahramani and Matthew J. Beal, "Variational inference for bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems*, Vol. 12 (2000) pp. 449–455.

[27] Michael E. Tipping and Christopher M. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation **11**, 443–482 (1999).

[28] Jalil Taghia, Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, Weidong Cai, and Vinod Menon, "Bayesian switching factor analysis for estimating time-varying functional connectivity in fmri," NeuroImage **155**, 271–290 (2017).

[29] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski, "Two-way analysis of high-dimensional collinear data," Data Mining and Knowledge Discovery **19**, 261–276 (2009).

[30] Shiwen Zhao, Chuan Gao, Sayan Mukherjee, and Barbara E Engelhardt, "Bayesian group factor analysis with structured sparsity," Journal of Machine Learning Research **17**, 1–47 (2016).

[31] Lurdes Y. T. Inoue, Mauricio Neira, Colleen Nelson, Martin Gleave, and Ruth Etzioni, "Cluster-based network model for time-course gene expression data," Biostatistics **8**, 507–525 (2007).

[32] Peter D Hoff, Adrian E Raftery, and Mark S Handcock, "Latent space approaches to social network analysis," Journal of the american Statistical association **97**, 1090–1098 (2002).

[33] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum, "Model-based clustering for social networks," Journal of the Royal Statistical Society A **170**, 301–354 (2007).

[34] Robert E. Kass and Adrian E. Raftery, "Bayes factors," Journal of the American Statistical Association **90**, 773–795 (1995).

[35] Jan Drugowitsch, "Variational bayesian inference for linear and logistic regression," arXiv , 1310.5438 (2013).

[36] Ignacio Alvarez, Jarad Niemi, and Matt Simpson, "Bayesian infernece for a covariance matrix," in *Annual conference on applied statistics in agriculture* (2014).

[37] Jaakko Luttinen, "Fast variational bayesian linear state-space model," in *European Conference on Machine Learning and Knowledge Discovery in Databases*, Vol. 8188 (2013) pp. 305–320.

[38] Christopher M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2007).

[39] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe, "Variational inference: A review for statisticians," Journal of the American Statistical Association **112**, 859–877 (2017).

[40] Michael Salter-Townshend and Thomas Brendan Murphy, "Variational bayesian inference for the latent position cluster model for network data," Computational Statistics & Data Analysis **57**, 661–671 (2013).

[41] Michael E. Tipping and Christopher M. Bishop, "Probabilistic principal component analysis," Journal of the Royal Statistical Society B **61**, 611–622 (1999).

[42] David Arthur and Sergei Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (2007) pp. 1027–1035.

[43] Alexander Strehl and Joydeep Ghosh, "Cluster ensembles: a knowledge reuse framework for combining multiple partitions," Journal of Machine Learning Research **3**, 583–617 (2002).

[44] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, "Inference and phase transitions in the detection of modules in sparse networks," Physical Review Letters **107**, 065701 (2011).

[45] Daniel J. Fenn, Mason A. Porter, Stacy Williams, Mark McDonald, Neil F. Johnson, and Nick S. Jones, "Temporal evolution of financial-market correlations," Phys. Rev. E **84**, 026109 (2011).

[46] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research **9**, 2579–2605 (2008).

[47] Leto Peel, Daniel B Larremore, and Aaron Clauset, "The ground truth about metadata and community detection in networks," Science Advances **3**, e1602548 (2017).

[48] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel, "World map of the köppen-geiger climate classification updated," Meteorologische Zeitschrift **15**, 259–263 (2006).

[49] Wladimir Köppen, "Die wärmezonen der erde, nach der dauer der heissen, gemässigten und kalten zeit und nach der wirkung der wärme auf die organische welt betrachtet," Meteorologische Zeitschrift **1**, 5–226 (1884).

[50] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment **2008**, P10008 (2008).

[51] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," Proceedings of the National Academy of Sciences **107**, 12755–12760 (2010), http://www.pnas.org/content/107/29/12755.full.pdf.

[52] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel, "Detectability thresholds and optimal algorithms for community structure in dynamic networks," Physical Review X **6**, 031005 (2016).

[53] Dane Taylor, Saray Shai, Natalie Stanley, and Peter J Mucha, "Enhanced detectability of community structure in multilayer networks through layer aggregation," Physical review letters **116**, 228301 (2016).

Appendix A: Exponential family distributions

For completeness and the convenience of readers with a non-Bayesian background, we provide definitions of some distributions.

1. Normal distribution

The univariate normal distribution with mean $\mu \in \mathbb{R}$ and precision $\tau > 0$ is denoted by $\text{Normal}(\mu, \tau^{-1})$ and has probability distribution

$$P(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(x-\mu)^2\right].$$

The multivariate normal distribution with mean vector $\mu \in \mathbb{R}^p$ and positive-definite precision matrix Λ is denoted by $\text{Normal}(\mu, \Lambda^{-1})$ and has probability distribution

$$P(x|\mu,\Lambda) = \sqrt{\frac{\det\Lambda}{(2\pi)^p}} \exp\left[-\frac{1}{2}(x-\mu)'\Lambda(x-\mu)\right].$$

2. Gamma and Wishart distributions

The Gamma distribution with shape parameter a and scale parameter b is denoted by $\text{Gamma}(a, b)$ and has probability distribution

$$P(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx),$$

where Γ is the Gamma function. The mean and variance of the Gamma distribution are

$$\langle x \rangle = \frac{a}{b} \tag{A1}$$

$$\text{var}x = \frac{a}{b^2}. \tag{A2}$$

The Wishart distribution $\text{Wishart}(\nu, W)$ is a multivariate generalisation of the Gamma distribution parametrised by the shape parameter $\nu > p-1$ and positive-definite scale parameter $W \in \mathbb{R}^{p\times p}$. It has probability distribution

$$P(x|\nu,W) = \frac{(\det W)^{\nu/2}}{2^{\nu p/2}\Gamma_p\left(\frac{\nu}{2}\right)} (\det x)^{(\nu-p-1)/2} \exp\left[-\frac{\text{tr}Wx}{2}\right],$$

where Γ_p is the multivariate Gamma function and $\text{tr}W$ denotes the trace of W. The parametrisation of the Wishart distribution is chosen to match the parametrisation of the Gamma distribution. Other texts may use $W \to W^{-1}$ instead. The mean and variance of the Wishart distribution are

$$\langle x \rangle = \nu V \tag{A3}$$

$$\text{var}x_{ij} = \nu\left(V_{ij}^2 + V_{ii}V_{jj}\right), \tag{A4}$$

where $V = W^{-1}$.

3. Dirichlet distribution

The Dirichlet distribution with concentration parameter $\alpha \in \mathbb{R}^p$ is denoted by $\text{Dirichlet}(\alpha)$ and has probability distribution

$$P(x|\alpha) = \frac{\Gamma(A)}{\prod_{i=1}^p \Gamma(\alpha_i)} \prod_{i=1}^p x_i^{\alpha_i - 1},$$

where $A = \sum_{i=1}^{p} \alpha_i$. The mean and variance of the Dirichlet distribution are

$$\langle x_i \rangle = \frac{\alpha_i}{A} \tag{A5}$$

$$\mathrm{var} x_i = \frac{\alpha_i (A - \alpha_i)}{A^2 (A+1)}. \tag{A6}$$

Appendix B: Update rules for variational inference

In this section, we derive the update rules for the posterior factors using the variational mean-field approximation. We start with the logarithm of the joint distribution of the time series observations and the model parameters

$$\log P(y, x, \tau, A, \mu, \Lambda, z, \lambda, \rho | a, b, \alpha, \beta, \gamma, \nu, W) \tag{B1}$$

$$= \log P(y|x, \tau, A) + \log P(x) + \log P(\tau|\alpha, \beta) + \log P(A|\mu, \Lambda, z) + \log P(\mu|\lambda) \tag{B2}$$

$$+ \log P(\Lambda|W, \nu) + \log P(\lambda|a, b) + \log P(z|\rho) + \log P(\rho|\gamma). \tag{B3}$$

The factors approximating the posterior for each model parameter are equal to the expectation of the log-joint distribution with respect to all other parameters [38]. Thus, we only need to consider terms that explicitly depend on the parameter of interest because all other terms can be absorbed into the normalisation constant of the factor.

Starting with the latent factors, we find

$$Q_{x_{tq}} \doteq -\frac{x_{tq} x_{tq}}{2} - \left\langle \frac{\tau_i}{2} \left(y_{ti} - A_{iq} x_{tq}\right) \left(y_{ti} - A_{ir} x_{tr}\right) \right\rangle_{\setminus x_{tq}},$$

where we have used the Einstein summation convention such that repeated indices that do not appear on both sides of the equation are summed over. The precision of the additive noise has posterior factor

$$Q_{\tau_i} \doteq \frac{T \log \tau_i}{2} - \frac{\tau_i}{2} \left\langle \left(y_{ti} - A_{iq} x_{tq}\right) \left(y_{ti} - A_{ir} x_{tr}\right) \right\rangle + (\alpha - 1) \log \tau_i - \beta \tau_i.$$

The posterior factor for the factor loadings is

$$Q_{A_i} \doteq -\left\langle \frac{\tau_i}{2} \left(y_{ti} - A_{iq} x_{tq}\right) \left(y_{ti} - A_{ir} x_{tr}\right) \right\rangle_{\setminus A_i} - \left\langle \frac{z_{ik}}{2} \left(A_{iq} - \mu_{kq}\right) \Lambda_{kqr} \left(A_{ir} - \mu_{kr}\right) \right\rangle_{\setminus A_i}.$$

The posterior factor for the centres of the groups is

$$Q_{\mu_k} \doteq -\left\langle \frac{z_{ik}}{2} \left(A_{iq} - \mu_{kq}\right) \Lambda_{kqr} \left(A_{ir} - \mu_{kr}\right) \right\rangle_{\setminus \mu_k} - \left\langle \frac{\lambda_{kq}}{2} \right\rangle \mu_{kq} \mu_{kq}.$$

The posterior factor for the precisions of the groups is

$$Q_{\Lambda_k} \doteq \left\langle \frac{z_{ik}}{2} \log \det \Lambda_k - \frac{z_{ik}}{2} \left(A_{iq} - \mu_{kq}\right) \Lambda_{kqr} \left(A_{ir} - \mu_{kr}\right) \right\rangle_{\setminus \Lambda_k} \tag{B4}$$

$$+ \frac{n - p - 1}{2} \log \det \Lambda_k - \frac{1}{2} W_{qr} \Lambda_{qr}. \tag{B5}$$

The posterior factor for the group assignments is

$$Q_{z_i} \doteq \frac{z_{ik}}{2} \left\langle \log \det \Lambda_k - \left(A_{iq} - \mu_{kq}\right) \Lambda_{kqr} \left(A_{ir} - \mu_{kr}\right) \right\rangle + z_{ik} \left\langle \log \rho_k \right\rangle.$$

The posterior factor for the group sizes is

$$Q_{\rho_k} \doteq \langle z_{ik} \rangle \log \rho_k + (\gamma - 1) \log \rho_k.$$

The posterior factor for the precision of the group centres is

$$Q_{\lambda_{kq}} \doteq \frac{\log \lambda_{kq}}{2} - \frac{\lambda_{kq}}{2} \langle \mu_{kq} \mu_{kq} \rangle + (a - 1) \log \lambda_{kq} - b \lambda_{kq}$$

$$\doteq \left(a - \frac{1}{2}\right) \log \lambda_{kq} - \left(b + \frac{\langle \mu_{kq} \mu_{kq} \rangle}{2}\right) \lambda_{kq}.$$