Some notes on the state-space LBM

Abstract

In response to the global coronavirus pandemic, we study co-clustering of multivariate time-series data as a way to simultaneously cluster both geographical regions and time periods after the outbreak of the pandemic. The resulting blocks of clusters, identified with a latent block model component, are integrated with an extended SIR model that takes both geographical clusters as well as clusters of time periods into account. We consider US data (or global or both?) and show how the novel latent block SIR model produce better prediction accuracy of the epidemic, and gives further understanding of how the virus is spreading in different geographical regions during different periods of time.

1 Background

The latent block model (Gérard Govaert and Nadif, 2003) is a commonly used model for co-clustering of large data matrices. It is a model-based approach that assumes that the rows and columns of the data matrix can be arranged according to latent row and column clusters. It has been extended to cover the case of counting data (Gérard Govaert and Nadif, 2010), continuous data (Nadif and Gerard Govaert, 2010), categorical data (Keribin et al., 2015), ordinal data (Jacques and Biernacki, 2018; Corneli, Charles Bouveyron, and Latouche, 2020), functional data (C. Bouveyron et al., 2018) and tensor data (Boutalbi, Labiod, and Nadif, 2020). In this work, the latent block model is extended in two ways: first by covering time-series data of proportions, and secondly by being integrated with the SIR model (Kermack and McKendrick, 1927), commonly used in the analysis of disease spread.

The class of SIR models are used to describe the spread of disease in a population. For a review of infective disease modeling, see Keeling and Rohani (2011). Compartmental models such as the SIR model has been used for modeling of infective disease for more than a 100 years (Ross, 1911), and have historically been used to e.g. understand disease transmission dynamics (Mills, Robins, and Lipsitch, 2004). With the 2020 coronavirus pandemic, attention has been focused on forecasting. Special attention has been given to key outbreak measures like peak intensity and the timing of such intensity. With accurate estimates of these quantities, policy makers have the opportunity to make informed decisions on allocation of resources, implementation of interventions and communication to the public (Chretien et al., 2014; Nsoesie, Mararthe, and Brownstein, 2013). For the simplest case of the SIR model, a population of susceptible cases $S(t)$ are exposed to a fraction $I(t)$ of the population that are infected at time point t. As a consequence, a part of the susceptible cases will be transferred to the

group of removed cases, denoted $R(t)$. The removed cases consist of both recovered and dead cases, and once a case enter this group there is no possibility to enter the group of susceptible and infected again. Following from this definition, $S(t) + I(t) + R(t) = 1$ for all $t \in [0, T]$ and $S(t), I(t), R(t) > 0$. This dynamic system is described by the following differential equations:

$$\frac{dS(t)}{dt} = -\rho S(t)I(t), \quad \frac{dI(t)}{dt} = \rho S(t)I(t) - \gamma I(t), \quad \frac{dR(t)}{dt} = \gamma I(t), \quad (1)$$

where $\rho > 0$ is the transmission rate of the disease, and $\gamma > 0$ is the rate of recovery. The model in (1) describes the transport of cases between the three groups of susceptible, infected and removed individuals. It is therefore often called a compartmental model (Ramsay and Hooker, 2017). Note that although the transmission rate ρ is constant, the rate of transport between the group of susceptible and infected cases does depend on time since the number of susceptible cases decreases as the number of infected increases. This follows from the fact that the total group of susceptible, infected and removed cases at all times equals the total population. Note also that the SIR model implicitly assumes that the cases are mixed freely with the entire population (Ramsay and Hooker, 2017).

2 The Extended SIR Model

Compartmental models like the SIR model are purely deterministic, but recently, several probabilistic versions of such models have been suggested. One of the most recent proposals is given in Osthus et al. (2017), where a state-space model motivated by the SIR model is presented, where inference and forecasting is conducted under a Bayesian setting. In Song et al. (2020), the model was extended to cover multivariate time-series, and a time-varying transmission rate. In the following, this model will be referred to as the extended SIR (eSIR) model.

To specify the state-space formulation of the eSIR model we introduce the following notation: $\mathbf{y} = (\mathbf{y}_i(t), i = 1, \ldots, n; t = 1, \ldots T)$ denotes the data matrix which is a multivariate time series. In the context of analyzing the 2020 coronavirus pandemic, $\mathbf{y}_i(t) = (y_i^I(t), y_i^R(t))^\top$ where $y_i(t)^I$ and $y_i(t)^R$ denote the proportion of infected and removed (recovered or dead) by the virus, respectively, at time point t. The index i denotes a geographical region, like for example country or state. Further, let $\boldsymbol{\theta} = (\theta_t^S, \theta_t^I, \theta_t^R)^\top$, where θ_t^S, θ_t^I and θ_t^R is the probability of a person being susceptible, infected and removed, respectively, at time point t. They thus satisfy $\theta_t^I + \theta_t^I + \theta_t^R = 1$ and $\theta_t^S, \theta_t^I, \theta_t^R > 0$ for all $t \in [0, T]$ (Osthus et al., 2017). It is assumed that $\boldsymbol{\theta}_{0:T} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ is a first-order Markov chain. This implies that $g(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:(t-1)}) = g(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \, \forall t \in [0 : T]$. Specifically, the following model for $\boldsymbol{\theta}$ is adopted:

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Omega}_1 \sim \text{Dirichlet}(\kappa f(\boldsymbol{\theta}_{t-1})),$$

where $\boldsymbol{\Omega}_1$ denotes the set of model parameters, κ scales the variance of the Dirichlet distribution, and the function $f(\cdot)$ is a 3-dimensional vector that sets the mean of the Dirichlet distribution. The form of the function $f(\cdot)$ will be presented in what follows. For the observed data $\mathbf{y}$, Song et al. (2020) make the

following distributional assumptions,

$$\begin{aligned} y_i^I(t)|\boldsymbol{\theta}, \boldsymbol{\Omega}_1 &\sim \text{Beta}(\lambda^I \theta_t^I, \lambda^I(1-\theta_t^I)) \\ y_i^R(t)|\boldsymbol{\theta}, \boldsymbol{\Omega}_1 &\sim \text{Beta}(\lambda^R \theta_t^R, \lambda^R(1-\theta_t^R) \end{aligned} \tag{2}$$

for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $\boldsymbol{\Omega}_1 = (\rho, \gamma, \boldsymbol{\theta}, \lambda^I, \lambda^R, \kappa)^\top$, where λ^I scales the variance in each respective distribution of $y_i^I(t)$ and $y_i^R(t)$.

The eSIR model thus considers a bivariate stochastic process $\{\boldsymbol{\theta}_t, \mathbf{y}_t\}$ that is modeled using a state-space model: $\boldsymbol{\theta}_t$ is the underlying, latent process that guides the observed data $\boldsymbol{y}_t = (y_i^I(t), y_i^R(t))$. This can be graphically summarized as

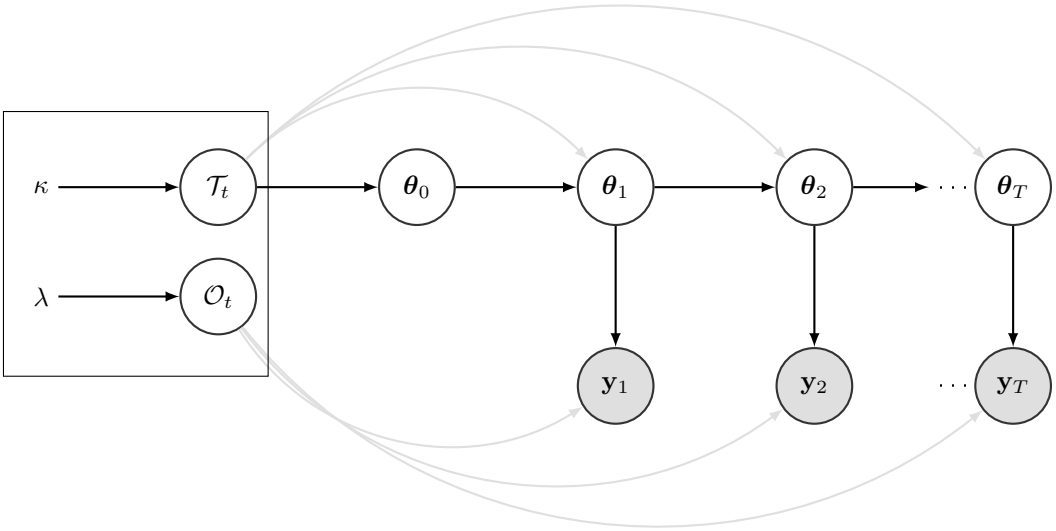


Figure 1: Visualization of the eSIR model.

Regarding the function $f(\cdot)$, it is the solution to the dynamic system

$$\frac{d\theta_t^S}{dt} = -\rho\pi(t)\theta_t^S\theta_t^I, \quad \frac{d\theta_t^I}{dt} = \rho\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I, \quad \frac{d\theta_t^R}{dt} = \gamma\theta_t^I, \tag{3}$$

where the term $\pi(t)$ is a transmission modifier equal to $\pi(t) = (1 - q^S(t))(1 - q^I(t))$, where $q^S(t)$ denotes the probability of a susceptible person being in-home isolation, and $q^I(t)$ the probability of an infected person being in-hospital quarantine. The term $\pi(t)$ therefore is a transmission modifier in the sense that it modifies the probability of a susceptible person getting in contact with an infected person. The chance of such a meeting to occur is to a great extent determined by which restrictions on social gatherings are in place. For example, if a geographical region does not impose a quarantine, $\pi(t) = 1$, and the dynamic system in (3) reduces to the classic formulation of the SIR model. As the rules of social distancing gets stricter, the transmission modifier $\pi(t)$ decreases, making

the overall transmission rate to decrease as well. In this work, the $\pi(t)$ term is allowed to differ from 1.

Since there are no explicit solutions available to (3), the so-called fourth-order Runge-Kutta approximation is implemented, meaning that

$$\begin{pmatrix} f(\theta^S_{t-1}) \\ f(\theta^I_{t-1}) \\ f(\theta^R_{t-1}) \end{pmatrix} = \begin{pmatrix} \theta^S_{t-1} + 1/6[k^{\theta^S_1}_{t-1} + 2k^{\theta^S_2}_{t-1} + 2k^{\theta^S_3}_{t-1} + k^{\theta^S_4}_{t-1}] \\ \theta^I_{t-1} + 1/6[k^{\theta^I_1}_{t-1} + 2k^{\theta^I_2}_{t-1} + 2k^{\theta^I_3}_{t-1} + k^{\theta^I_4}_{t-1}] \\ \theta^R_{t-1} + 1/6[k^{\theta^R_1}_{t-1} + 2k^{\theta^R_2}_{t-1} + 2k^{\theta^R_3}_{t-1} + k^{\theta^R_4}_{t-1}] \end{pmatrix}$$

where

$$\begin{aligned} k^{\theta^S_1}_{t-1} &= -\rho\pi(t-1)\theta^S_{t-1}\theta^I_{t-1}, \\ k^{\theta^S_2}_{t-1} &= -\rho\pi(t-1)[\theta^S_{t-1} + 0.5k^{\theta^S_1}_{t-1}][\theta^I_{t-1} + 0.5k^{\theta^I_1}_{t-1}], \\ k^{\theta^S_3}_{t-1} &= \rho\pi(t-1)[\theta^S_{t-1} + 0.5k^{\theta^S_2}_{t-1}][\theta^I_{t-1} + 0.5k^{\theta^I_2}_{t-1}], \\ k^{\theta^S_4}_{t-1} &= \rho\pi(t-1)[\theta^S_{t-1} + k^{\theta^S_3}_{t-1}][\theta^I_{t-1} + k^{\theta^I_3}_{t-1}], \end{aligned}$$

$$\begin{aligned} k^{\theta^I_1}_{t-1} &= \rho\pi(t-1)\theta^S_{t-1}\theta^I_{t-1} - \gamma\theta^I_{t-1}, \\ k^{\theta^I_2}_{t-1} &= \rho\pi(t-1)[\theta^S_{t-1} + 0.5k^{\theta^S_1}_{t-1}][\theta^I_{t-1} + 0.5k^{\theta^I_1}_{t-1}] - \gamma[\theta^I_{t-1} + 0.5k^{\theta^I_1}_{t-1}], \\ k^{\theta^I_3}_{t-1} &= \rho\pi(t-1)[\theta^S_{t-1} + 0.5k^{\theta^S_2}_{t-1}][\theta^I_{t-1} + 0.5k^{\theta^I_2}_{t-1}] - \gamma[\theta^I_{t-1} + 0.5k^{\theta^I_2}_{t-1}], \\ k^{\theta^I_4}_{t-1} &= \rho\pi(t-1)[\theta^S_{t-1} + k^{\theta^S_3}_{t-1}][\theta^I_{t-1} + k^{\theta^I_3}_{t-1}] - \gamma[\theta^I_{t-1} + k^{\theta^I_3}_{t-1}], \end{aligned}$$

and

$$\begin{aligned} k^{\theta^R_1}_{t-1} &= \gamma\theta^I_{t-1}, \\ k^{\theta^R_2}_{t-1} &= \gamma[\theta^I_{t-1} + 0.5k^{\theta^I_1}_{t-1}], \\ k^{\theta^R_3}_{t-1} &= \gamma[\theta^I_{t-1} + 0.5k^{\theta^I_2}_{t-1}], \\ k^{\theta^R_4}_{t-1} &= \gamma[\theta^I_{t-1} + k^{\theta^I_3}_{t-1}]. \end{aligned}$$

3 The State-Space Latent Block Model

Both the SIR model and the eSIR model assume that the population of cases are freely mixed in the population, although in many real situations the mixing of a specific group of cases might have geographical limitations. Moreover, although the eSIR allows for a varying transmission rate ρ, the modifier function $\phi(t)$ needs to be defined by the modeler. With a pandemic like the 2020 coronavirus, the modifier function would furthermore differ between different geographical regions. With the aim to increase predictive accuracy, we address these issues by incorporating into the eSIR model a latent block model component, presented in the following. The general idea is to simultaneously identify row clusters of geographical regions and column clusters of time points in the data matrix $\mathbf{y}$, and then re-define the eSIR model on these identified block clusters. In this way we will both be able to make forecasts on geographical regions identified by the row clusters, and estimate the transmission modifier $\pi(t)$ with data by setting it according to the estimated column clusters.

With n geographical regions measured on T time points, the data matrix $\mathbf{y}$ that we wish to co-cluster equals

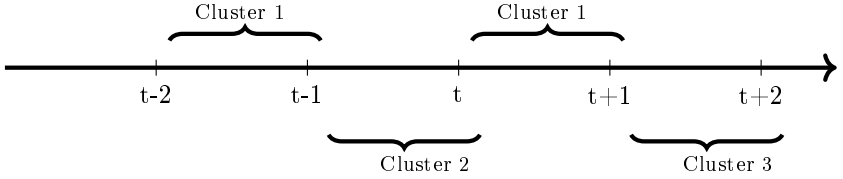
$$\mathbf{y} = \begin{bmatrix} (y_1^I(1), y_1^R(1)) & (y_1^I(2), y_1^R(2)) & \dots & (y_1^I(T), y_1^R(T)) \\ (y_2^I(1), y_2^R(1)) & (y_2^I(2), y_2^R(2)) & \dots & (y_2^I(T), y_2^R(T)) \\ \vdots & \vdots & \dots & \vdots \\ (y_n^I(1), y_n^R(1)) & (y_n^I(2), y_n^R(2)) & \dots & (y_n^I(T), y_n^R(T)) \end{bmatrix}$$

Following the latent block model, we assume that there is a partition (Z, W) of the data matrix $\mathbf{y}$, where $Z = (z_{ik};, i = 1, \dots, n,\ k = 1, \dots, K)$ represents the partitioning into K clusters on the n rows and $W = (w_{tl}; t = 1, \dots, T,\ l = 1, \dots, L)$ represents the partitioning into L clusters on the T measured time-points (columns). In other words, Z_{ik}, $k = 1, ..., K$ and W_{tl}, $l = 1, ..., L$ are binary matrices for which $Z_{ik} = 1$ if case i belongs to row cluster k and 0 otherwise, and $W_{tl} = 1$ if time point t belongs to column cluster l and 0 otherwise. The random matrices Z and W therefore are of dimension $n \times K$ and $T \times L$, respectively.

Co-clustering will yield subgroups, called blocks, such that $Z_{ik}W_{tl} = 1$. Each element $\mathbf{y}_i(t)$ in $\mathbf{y}$ belongs to a block which is generated by a probability distribution. In this study it is assumed that $y_i^I(t)$ and $y_i^R(t)$ follow Beta distributions, meaning that these block distributions are given by the distributions specified by Equation 2. Specifically, we assume that for a geographical region i belonging to row cluster k and a time-point t belonging to column cluster l, the block $Z_{ik}W_{jl}$ is generated by the following distributions:

$$\begin{aligned} y_i^I | Z_{ik}W_{jl} = 1, \mathbf{\Omega}_1 &\sim \text{Beta}(\lambda^I \theta_{kl}^I, \lambda^I (1 - \theta_{kl}^I) \\ y_i^R | Z_{ik}W_{jl} = 1, \mathbf{\Omega}_1 &\sim \text{Beta}(\lambda^R \theta_{kl}^R, \lambda^R (1 - \theta_{kl}^R) \end{aligned} \tag{4}$$

where θ_{kl}^I and θ_{kl}^R denote the probabilities of a case in row cluster k and column cluster l being infected and removed, respectively. Regarding the state process, we are going to deviate from the eSIR model and the Dirichlet distribution. The reason is that θ_{kl} is now guiding the observed data, and the time clusters are allowed to re-appear in the time interval. To illustrate that idea, consider Figure 3. We have, by co-clustering the data matrix using the LBM, ended up with three time (column) clusters. To illustrate how these clusters spread out across time, we



According to the figure, the first time (column) cluster equals the time period between $t-2$ and $t-1$. The next period of time, between $t-1$ and t belongs to cluster 2, and after that we are back at cluster 1 before entering cluster 3. The LBM offers us the flexibility to return to clusters as we progress in time. This property seems well-suited when modeling the coronavirus pandemic as we for example currently are seeing a widespread second wave hit many countries.

To model the transition between time clusters, we need to estimate the transition matrix $A = [a_{l,l'}]$, $[a_{l,l'}] = P(\boldsymbol{\theta} = \boldsymbol{\theta}_{kl'} | \boldsymbol{\theta} = \boldsymbol{\theta}_{kl})$

We thus assume that the latent process $\boldsymbol{\theta}(t)$ guides the observed data in each block cluster, that in turn is distributed according to a Beta distribution.

It is assumed that Z and W are independent from each other and that the random variables $\mathbf{y}$ are independent conditional on Z and W. With this formulation of the model we further assume that there is a partition of geographical regions that are homogeneous in terms of proportion infected $y^I(t)$ and removed $y^R(t)$, and that there is a partitioning of the time points t that form homogeneous blocks.

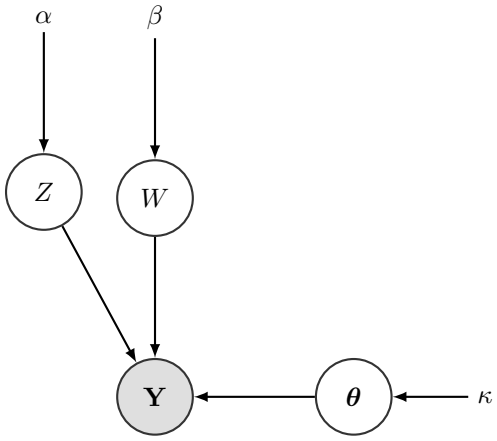
Now let $\alpha_k = P(Z_{ik} = 1)$ and $\beta_l = P(W_{jl} = 1)$ denote the respective row and column mixing proportions such that they both sum to 1 and $p(z;\theta) = \prod_{ik} \alpha_k^{z_{ik}}$ and $p(w;\theta) = \prod_{ik} \beta_l^{w_{jl}}$. Under the assumption of Z and W being independent, and by letting $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of all possible partitions of Z and W, the likelihood of the LBM equals

$$L(\boldsymbol{\Omega}_2) = \sum_{(z,w)\in\mathcal{Z},\mathcal{W}} \prod_{i,g} \alpha^{z_{ig}} \prod_{j,l} \beta^{w_{jl}} \prod_{i,j,k,l} \varphi(\mathbf{y}_{ij};\omega_{kl})^{z_{ig}w_{jl}},$$

where ω_{kl} represents the parameter of φ for the kl block. The log-likelihood equals

$$\log L(\boldsymbol{\Omega}_2) = \sum_{i,k} z_{ik} \log \alpha_k \sum_{j,l} w_{jl} \log \beta_l \sum_{i,j,k,l} z_{ik} w_{jl} \log \varphi(\mathbf{y}_{ij};\omega_{kl}),$$

The state-space LBM can now graphically be represented as



4 Estimation

Since there are two model components of the eSIR LBM, we repeat the total set of parameters that needs to be estimated. The unknown parameters from the eSIR model component of the likelihood thus equals $\boldsymbol{\Omega}_1 = (\rho, \gamma, \boldsymbol{\theta}, \lambda, \kappa)$ and the LBM model component of the likelihood equals $\boldsymbol{\Omega}_2 = (\alpha, \beta, \omega)$. The total set of parameters to be estimated thus equals $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 = (\rho, \gamma, \boldsymbol{\theta}, \lambda, \kappa, \alpha, \beta, \omega)$.

For the estimation of the eSIR LBM model, we will assume that $\varphi(\mathbf{y}_{ij};\omega_{kl})$ follows a Dirichlet distribution:

$$\varphi(\mathbf{y}_{ij};\omega_{kl}) = D(\omega_{kl})^{-1} \prod_{j=1}^{d+1} y_{ij}^{\omega_{kl}-1}$$

So should we model $\varphi(\cdot)$ as a bivariate beta distribution (meaning Dirichlet distribution)? Regarding this, see the paper "Time Series of Continuous Proportions", by Grunwald, Raftery and Guttorp (1993), where they model the time series of proportions using the Dirichlet distribution.

There is a paper ("Estimation and selection for the latent block model on categorical data" by Keribin et al.) that implements the LBM for multinomial data that in the estimation of the model sets prior distributions for the mixing proportions as well as the parameter that governs the Y distribution. This would in a sense be similar to our case, since the eSIR model imposes a Dirichlet prior on the $\boldsymbol{\theta}$ parameter. If we would further impose Dirichlet priors on the mixing proportions, would we be able to do something similar as in Keribin et al.?

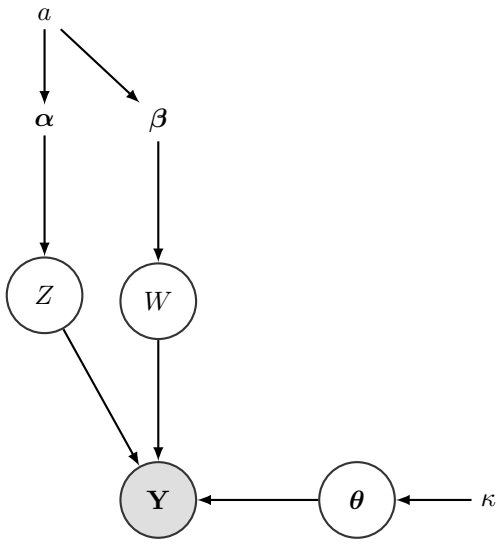
So specifically, following (Keribin et al., 2015) we can consider proper and non-informative priors for α and β as

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{Dirichlet}(a, \ldots, a) \\ \boldsymbol{\beta} &\sim \text{Dirichlet}(a, \ldots, a) \end{aligned} \tag{5}$$

In (Keribin et al., 2015) they consider a very similar general modeling structure and estimate the model parameters $\boldsymbol{\tau}$ by maximizing the posterior density $p(\boldsymbol{\tau}|\boldsymbol{y})$, which leads to the Maximum A Posteriori (MAP) estimator:

$$\hat{\boldsymbol{\tau}}_{MAP} = \underset{\boldsymbol{\tau}}{\text{argmax}}\, p(\boldsymbol{\tau}|\boldsymbol{y}) \tag{6}$$

We would thus be able to graphically represent the model as



References

Boutalbi, Rafika, Lazhar Labiod, and Mohamed Nadif (2020). "Tensor latent block model for co-clustering". In: *International Journal of Data Science and Analytics*, pp. 1–15.

Bouveyron, C. et al. (2018). "The Functional Latent Block Model for theCo-Clustering of Electricity Consumption Curves". In: *Journal of the Royal Statistical Society, Series C Applied Statistics*(67), pp. 897–915.

Chretien, Jean-Paul et al. (2014). "Influenza forecasting in human populations: a scoping review". In: *PloS one* 9(4), e94130.

Corneli, Marco, Charles Bouveyron, and Pierre Latouche (2020). "Co-clustering of ordinal data via latent continuous random variables and not missing at random entries". In: *Journal of Computational and Graphical Statistics*, pp. 1–15.

Govaert, Gérard and Mohamed Nadif (2003). "Clustering with block mixture models". In: *Pattern Recognition* 36(2), pp. 463–473.

Govaert, Gérard and Mohamed Nadif (2010). "Latent block model for contingency table". In: *Communications in Statistics—Theory and Methods* 39(3), pp. 416–425.

Jacques, Julien and Christophe Biernacki (2018). "Model-based co-clustering for ordinal data". In: *Computational Statistics & Data Analysis* 123, pp. 101–115.

Keeling, Matt J and Pejman Rohani (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Keribin, Christine et al. (2015). "Estimation and selection for the latent block model on categorical data". In: *Statistics and Computing* 25(6), pp. 1201–1216.

Kermack, William Ogilvy and Anderson G McKendrick (1927). "A contribution to the mathematical theory of epidemics". In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115(772), pp. 700–721.

Mills, Christina E, James M Robins, and Marc Lipsitch (2004). "Transmissibility of 1918 pandemic influenza". In: *Nature* 432(7019), pp. 904–906.

Nadif, Mohamed and Gerard Govaert (2010). "Model-based co-clustering for continuous data". In: *2010 Ninth international conference on machine learning and applications*. IEEE, pp. 175–180.

Nsoesie, Elaine, Madhav Mararthe, and John Brownstein (2013). "Forecasting peaks of seasonal influenza epidemics". In: *PLoS currents* 5.

Osthus, Dave et al. (2017). "Forecasting seasonal influenza with a state-space SIR model". In: *The annals of applied statistics* 11(1), p. 202.

Ramsay, James and Giles Hooker (2017). "Dynamic data analysis". In:

Ross, Ronald (1911). *Some quantitative studies in epidemiology*.

Song, Peter X et al. (2020). "An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China". In: *MedRxiv*.