Some notes on state-space LBM

Gabriel Wallin

October 5, 2020

1 SIR model

To specify the State-Space Latent Block Model (SS-LBM), we introduce the following notation: $\mathbf{y} = (\mathbf{y}_{ij}, i = 1, \ldots, n; j = 1, \ldots d)$ denotes the data matrix. Each element in $\mathbf{y}_{ij}$ is a multivariate time series: $\mathbf{y}_{ij} = (y_{ij}^1(t), \ldots, y_{ij}^S(t))$ where $t \in [0, T]$. In the context of analyzing the 2020 coronavirus pandemic, $\mathbf{y}_{ij} = (y_{ij}^I(t), y_{ij}^R(t))^\top$ where $y_{ij}^I(t)$ and $y_{ij}^R(t)$ denote the proportion of infected and removed (recovered or dead) by the virus, respectively, at time point t. 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. We will assume that $\boldsymbol{\theta}_{0:T} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ is a first-order Markov chain in the same spirit as (Osthus et al. 2017) and (Song et al. 2020). 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, we assume the following model for $\boldsymbol{\theta}$:

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

where 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 function f is the solution to the following dynamic system:

$$\frac{d\theta_t^S}{dt} = -\rho\theta_t^S\theta_t^I, \quad \frac{d\theta_t^I}{dt} = \rho\theta_t^S\theta_t^I - \pi\theta_t^I, \quad \frac{d\theta_t^R}{dt} = \pi\theta_t^I, \tag{1}$$

where $\rho > 0$ is the transmission rate of the disease and $\pi > 0$ is the rate of recovery. This system of non-linear ordinary differential equations is within the field of epidemiology known as the SIR model (Kermack and McKendrick 1927). Since there are no explicit solutions available to (1), the so-called fourth-order Runge-Kutta approximation is implemented, meaning that

$$\begin{pmatrix} f(\theta_{t-1}^S) \\ f(\theta_{t-1}^I) \\ f(\theta_{t-1}^R) \end{pmatrix} = \begin{pmatrix} \theta_{t-1}^S + 1/6[k_{t-1}^{\theta_1^S} + 2k_{t-1}^{\theta_2^S} + 2k_{t-1}^{\theta_3^S} + k_{t-1}^{\theta_4^S}] \\ \theta_{t-1}^I + 1/6[k_{t-1}^{\theta_1^I} + 2k_{t-1}^{\theta_2^I} + 2k_{t-1}^{\theta_3^I} + k_{t-1}^{\theta_4^I}] \\ \theta_{t-1}^R + 1/6[k_{t-1}^{\theta_1^R} + 2k_{t-1}^{\theta_2^R} + 2k_{t-1}^{\theta_3^R} + k_{t-1}^{\theta_4^R}] \end{pmatrix}$$

where

$$\begin{aligned}
k_{t-1}^{\theta_1^S} &= -\rho\theta_{t-1}^S\theta_{t-1}^I, \\
k_{t-1}^{\theta_2^S} &= -\rho[\theta_{t-1}^S + 0.5k_{t-1}^{\theta_1^S}][\theta_{t-1}^I + 0.5k_{t-1}^{\theta_1^I}], \\
k_{t-1}^{\theta_3^S} &= \rho[\theta_{t-1}^S + 0.5k_{t-1}^{\theta_2^S}][\theta_{t-1}^I + 0.5k_{t-1}^{\theta_2^I}], \\
k_{t-1}^{\theta_4^S} &= \rho[\theta_{t-1}^S + k_{t-1}^{\theta_3^S}][\theta_{t-1}^I + k_{t-1}^{\theta_3^I}],
\end{aligned}$$

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

and

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

Lastly, for the data $\mathbf{y}$ we follow (Song et al. 2020) and assume the following state-space model,

$$
\begin{aligned}
y^I_{ij}|\boldsymbol{\theta}, \boldsymbol{\Omega}_1 &\sim \mathrm{Beta}(\lambda^I\theta^I_t, \lambda^I(1-\theta^I_t))\\
y^R_{ij}|\boldsymbol{\theta}, \boldsymbol{\Omega}_1 &\sim \mathrm{Beta}(\lambda^R\theta^R_t, \lambda^R(1-\theta^R_t)
\end{aligned}
\tag{2}
$$

where $\boldsymbol{\Omega}_1 = (\rho, \pi, \theta, \lambda, \kappa)^\top$.

The SIR model can thus be graphically summarized as

$$
\begin{tikzcd}
\text{Markov process:} & \boldsymbol{\theta}_0 & \boldsymbol{\theta}_1 \arrow[d] & \boldsymbol{\theta}_2 \arrow[d] & \cdots & \boldsymbol{\theta}_{T-1} \arrow[d] \\
\text{Observations:} & \mathbf{y}_0 & \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{T-1}
\end{tikzcd}
$$

2 Latent Block Model

The latent block model (LBM; Govaert and Nadif 2003) assumes a partitioning of the rows and columns of the data matrix $\mathbf{y}$ into blocks. Specifically, there is a partition (Z, W) for which Z is partitioned into K clusters on the n rows and W is partitioned into L clusters on the d columns. In other words, Z_{ik}, $k = 1, ..., K$ and W_{jl}, $l = 1, ..., L$ are binary matrices for which $Z_{ik} = 1$ if case i belongs to row cluster k and 0 otherwise, and $W_{jl} = 1$ if feature j belongs to column cluster l and 0 otherwise. The matrices Z and W therefore are of dimension $n \times K$ and $d \times L$, respectively. Co-clustering will yield subgroups, called blocks, such that $z_{ik}w_{jl} = 1$. Each element y_{ij} in $\mathbf{y}$ belongs to a block which is generated by a probability distribution. Since it is assumed that $y^I_{ij}(t)$ and $y^R_{ij}(t)$ follow Beta distributions, these block distributions are given by the distributions specified by Equation 2. 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.

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 $p(z; \theta) = \prod_{ik} \alpha_k^{z_{ik}}$

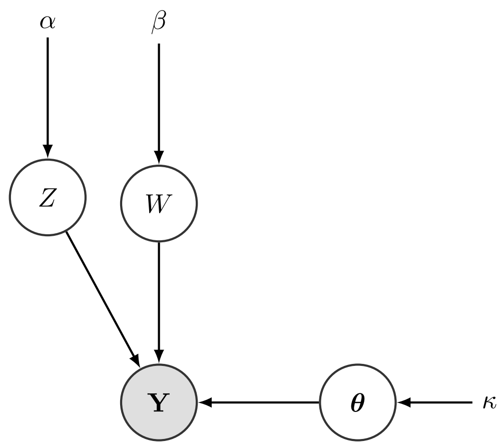
and $p(w;\theta) = \prod_{ik} \beta_l^{w_{jl}}$. Under the assumption of Z and W being independent, and by letting $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of all possible partitions of Z and W, the likelihood of the LBM equals

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

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



3 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 $\boldsymbol{\Omega}_1 = (\rho, \pi, \boldsymbol{\theta}, \lambda, \kappa)$ 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 $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 = (\rho, \pi, \boldsymbol{\theta}, \lambda, \kappa, \alpha, \beta, \omega)$.

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.?

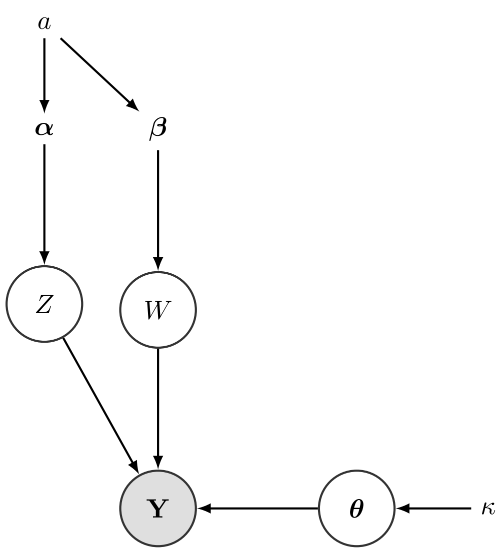
So specifically, following (Keribin et al. 2015) we can consider proper and non-informative priors for α and β as

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{Dirichlet}(a, \ldots, a) \\ \boldsymbol{\beta} &\sim \text{Dirichlet}(a, \ldots, a) \end{aligned} \tag{3}$$

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:

$$\hat{\boldsymbol{\tau}}_{MAP} = \underset{\boldsymbol{\tau}}{\operatorname{argmax}}\, p(\boldsymbol{\tau}|\boldsymbol{y}) \tag{4}$$

We would thus be able to graphically represent the model as



References

Govaert, Gérard and Mohamed Nadif (2003). "Clustering with block mixture models". In: *Pattern Recognition* 36.2, pp. 463–473.

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.

Osthus, Dave et al. (2017). "Forecasting seasonal influenza with a state-space SIR model". In: *The annals of applied statistics* 11.1, p. 202.

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