Summary of the state-space latent block model

1 Extending the SIR model for clustered data

Consider the data matrix

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

where $y_i^I(t)$ and $y_i^R(t)$, $i = 1, \dots, n$ and $t = 1, \dots, T$, denote the proportions of infected and removed in geographical region i at time point t. For each discrete time point t, we thus record $\mathbf{y}_i(t) = (y_i^I(t), y_i^R(t))^\top$ for geographical region i. The dimension of the data matrix is thus $n \times T$.

We assume that for each $\mathbf{y}_i$ there is an associated value of an unobserved label z such that $z_{ik} = 1$ if $\mathbf{y}_i$ belongs to the kth cluster. It follows that the observations come from K distinct populations, hereinafter referred to as components. Each component has its own distribution but which unit that belongs to which component is unknown. We address the problem of identifying these hidden clusters with a model-based approach. Assuming that the component distributions belong to the same parametric family $\varphi(\mathbf{y}; \phi_k)$ so that the component only differ by a parameter value, the mixing density equals

$$f(\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \varphi(\mathbf{y}; \phi_k), \tag{1}$$

where ϕ_k is the parameter vector for the kth component, and $\alpha_1, \dots, \alpha_K$ are the mixing weights, meaning that $P(z_{ik} = 1) = \alpha_k$. They fulfill $\alpha_k > 0$ and $\sum_k \alpha_k = 1$. In Equation 1, $\varphi(\cdot; \phi_k)$ thus denotes the density of the kth component given its parameter value ϕ_k.

The model and its assumptions are summarized as follows:

1. $\mathbf{z} = (z_{ik}; i = 1, \dots, n; k = 1, \dots, K)$ represents the clustering of rows into K groups, where row i belongs to cluster k if $z_{ik} = 1$. The group indicator of row i is denoted $\mathbf{z}_i$. These row labels are independent, and follow a multinomial distribution:

$$\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha} = (\alpha_1, \dots \alpha_K)).$$

2. Conditional on the row labels, the observed data $\mathbf{y}$ are independent and their conditional distribution follows a Dirichlet distribution with parameter only depending on the given cluster,

$$\mathbf{y}_i(t) | \{z_{ik} = 1\}, \phi_k \sim \mathcal{D}(\boldsymbol{\theta}_k(t)), \tag{2}$$

where $\boldsymbol{\theta}_k(t) = (\theta_k^S(t), \theta_k^I(t), \theta_k^R(t))$ denotes the probability of an individual being susceptible, infected and removed (recovered or died), respectively, in cluster k and at time point t.

3. For each cluster k, it is assumed that the model proportions $\boldsymbol{\theta}_k(0:T) = (\boldsymbol{\theta}_k(0), \boldsymbol{\theta}_k(1), \ldots, \boldsymbol{\theta}_k(T))$ form a first-order Markov chain. This implies that $g(\boldsymbol{\theta}_k(t)|\boldsymbol{\theta}_k(0:(t-1))) = g(\boldsymbol{\theta}_k(t)|\boldsymbol{\theta}_k(t-1))\, \forall t \in [0:T]$. It is furthermore assumed that $\boldsymbol{\theta}_k$ are independent, random vectors following a Dirichlet distribution,

$$\boldsymbol{\theta}_k(t)|\boldsymbol{\theta}_k(t-1), \phi_k \sim \mathcal{D}(\kappa f(\boldsymbol{\theta}_k(t-1))). \tag{3}$$

The term κ controls the variance, $\phi_k = (\boldsymbol{\alpha}, \boldsymbol{\theta}_k, \rho_k, \gamma_k, \kappa)$, and the function $f(\cdot) \in \mathbb{R}^3$ is the solution to the following system of nonlinear differential equations,

$$\begin{aligned} \frac{d\theta_k^S(t)}{dt} &= -\rho_k \theta_k^S(t)\theta_k^I(t), \\ \frac{d\theta_k^I(t)}{dt} &= \rho_k \theta_k^S(t)\theta_k^I(t) - \gamma_k \theta_k^I(t), \\ \frac{d\theta_k^R(t)}{dt} &= \gamma_k \theta_k^I(t), \end{aligned} \tag{4}$$

where ρ_k is the transmission rate in cluster k, and γ_k is the recovery rate in cluster k. In this way, we have a transmission and recovery rate that depends on geographical region. The so-called Runge-Kutta approximation can be implemented to solve the system in Equation 4.

2 Estimation

Following from the assumptions 1-3 above, the complete data likelihood of the proposed model equals

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}; \phi_k) = p(\mathbf{z}; \phi_k)p(\boldsymbol{\theta}; \phi_k)p(\mathbf{y}|\mathbf{z}; \phi_k) = \prod_{i,k} \alpha_k^{z_{ik}} \prod_{i,k} g(\boldsymbol{\theta}_k; \phi_k)^{z_{ik}} \prod_{i,k} \varphi(\mathbf{y}_i; \phi_k)^{z_{ik}} \tag{5}$$

where $g(\cdot)$ is specified by Equation 3 and $\varphi(\cdot)$ is specified by Equation 2.

Since the labels are unobserved, the observed likelihood, obtained by marginalizing over all label configurations, equals

$$p(\mathbf{y}; \phi_k) = \sum_{\mathbf{z} \in \mathcal{Z}} \left(\prod_{i,k} \alpha_k^{z_{ik}} \prod_{i,k} g(\boldsymbol{\theta}_k; \phi_k)^{z_{ik}} \prod_{i,k} \varphi(\mathbf{y}_i; \phi_k)^{z_{ik}} \right)$$

3 Questions/Things I'm thinking about

1. If we consider the suggested model by looking at Equations 2 and 3, is the time dependence clear this time? And if we look at the likelihood as stated in 5, is the time dependence still clear? I feel fairly confident that the time dependence is incorporated if we look at Equations 2 and 3, but less sure about the likelihood.

2. I have included the $\boldsymbol{\theta}$ component into the likelihood. Up until now I have left it out of the likelihood but it seems reasonable to that it should be included.

3. I am not sure how/if the EM algorithm (or some modified version of it) would work to estimate this model, given that we have two unobserved parts, z and $\boldsymbol{\theta}$. This is something that I will continue to think about until our meeting.

4. I am trying to implement the clustering idea presented here into the estimation framework suggested in the eSIR paper (MCMC using Gibbs sampling). I will hopefully have something to show you regarding this on Monday.