Approximating general SIS using non-negative matrix factorization

E. Cator, H. Don and P. Van Mieghem

September 27, 2016

Abstract

In this paper we consider the SIS model with general infection rate matrix $A$. Using a non-negative matrix factorization to approximate $A$, we are able to identify when a metastable state can be expected, and that the metastable distribution, under certain conditions, will be a normal distribution with known expectation and covariance. Also, we are able to model the behaviour of the model when starting outside of the metastable state, by approximating the behaviour by a standard linear SDE in high enough dimensions.

1 Introduction

We will study general SIS, where we have $n$ subjects (or nodes) that infect each other at given rates, and each subject can heal at its own rate. We introduce the non-negative matrix $A$ such that $A_{ij}$ is the rate at which subject $i$, when infected, tries to infect subject $j$. $A$ is not necessarily symmetric. Furthermore, we define the vector $\Delta$ such that $\Delta_i$ is the rate at which the infected subject $i$ heals.

Our proposal to study this general SIS process is twofold: first we approximate the matrix $A$ through non-negative matrix factorization (NMF), i.e., we approximate $A$ by $W^T H$, where $W$ and $H$ are $k \times n$ matrices (where in general $k < n$). We do this sparse approximation in such a way that we best approximate the off-diagonal elements of $A$, since the diagonal elements carry no information.

The second step in our approach is to study the behaviour of SIS whenever the infection matrix is of the form $W^T H$. It is well known that NMF can be used to perform a clustering of the $n$ subjects. We intend to use this clustering idea to translate the SIS to a solvable stochastic diffusion process, which is amenable not only to the analysis of the stationary, meta-stable, configuration, but even to the analysis of the time-evolution of SIS.

2 Non-negative matrix factorization

In recent years a lot of attention has been given to MNF. Several good algorithms have been developed to find the matrices $W$ and $H$ in $\mathbb{R}^{k \times n}$ that minimise given objective functions such as the Frobenius norm $\|A - W^T H\|$. A particularity of our specific problem is that the diagonal of $A$ is irrelevant. This may lead to different objective functions and possibly adapted algorithms for finding $W$ and $H$. In the case where $A$ is symmetric, it is common practice to choose $H = W$, but we will not do so in general.
3 SIS when $A = W^T H$

Suppose $A = W^T H$, and denote $W = (W_1 | \cdots | W_n)$ and $H = (H_1 | \cdots | H_n)$, where $W_i, H_i \in \mathbb{R}^k$. The infection rate with which node $i$ infects node $j$ is given by the inner product of $W_i$ and $H_j$:

$$A_{ij} = W_i^T H_j.$$ 

We therefore call $W_i$ the infectiousness of node $i$, and $H_j$ the susceptibility of node $j$. The final characteristic of a node is its healing rate $\Delta_i$.

So each node is characterized by the vector $Z_i \in \mathbb{R}^{2k+1}$, with

$$Z_i = \begin{pmatrix} \sqrt{n}W_i \\ \sqrt{n}H_i \\ \Delta_i \end{pmatrix}.$$ 

The scaling of the two components $W$ and $H$ is useful later on, as we will see. If two nodes $i$ and $j$ have the same "Z-vector" (so $Z_i = Z_j$), then these two nodes become indistinguishable in the SIS process. This means that the reduced process where we only remember how many of $i$ and $j$ are sick (0, 1 or 2) and not who is sick, is still a Markov process. If $n >> k$, then we have a lot of Z-vectors in $\mathbb{R}^{2k+1}_+$, and therefore a lot of nodes that are almost indistinguishable. We will choose $r$ clusters $C_j \subset \{1, \ldots, n\}$ such that

$$\bigcup_{j=1}^r C_j = \{1, \ldots, n\} \text{ and } j \neq l \implies C_j \cap C_l = \emptyset.$$ 

Then we define

$$n_j = \# C_j \text{ and } Y_j = \frac{1}{n_j} \sum_{i \in C_j} Z_i.$$ 

In other words, $Y_j$ equals the average of all Z-vectors in cluster $j$, and we will call it the cluster center of $C_j$. Of course we want to choose the clusters in such a way that almost all nodes have their Z-vector reasonably close to their corresponding $Y_j$. Also, we have to make a choice for $r$, the number of clusters: we will get back to this later. For now, the intuition we should have is that $r$ is big enough such that the cluster centers are a reasonable approximation for enough nodes, and not too big, so that the number of nodes in each cluster is large enough.

Once we have chosen our clusters, we will study a new SIS model that approximates the model with our original $A$. We do this by replacing the Z-vector of each node by its corresponding cluster center $Y_j$. In other words, define

$$\tilde{Z}_i = \sum_{j=1}^r 1_{\{i \in C_j\}} Y_j.$$ 

In this model, all nodes belonging to the same cluster have become indistinguishable, which mean that when we define $N = (N_1, \ldots, N_r)$ as the vector containing the number of infected nodes in each cluster, we will have that $N$ is a Markov process! We will study this Markov process in more detail.
3.1 The Markov process $N = (N_1, \ldots, N_r)$

We introduce the following notation: we split up any vector $y \in \mathbb{R}^{2k+1}$ such that

$$y = \begin{pmatrix} y_w \\ y_h \\ y_\delta \end{pmatrix},$$

where $y_w \in \mathbb{R}^k$ corresponds to the infectiousness, $y_h \in \mathbb{R}^k$ corresponds to the susceptibility and $y_\delta$ corresponds to the healing rate. Then we can determine the generator of the process $N$ by giving the possible transfer rates:

$$N_j \rightarrow N_j + 1 \text{ at rate } \frac{1}{n} \sum_{l=1}^{r} (n_j - N_j) N_l Y_{w,l}^T Y_{h,j}. \quad (1)$$

$$N_j \rightarrow N_j - 1 \text{ at rate } N_j Y_{\delta,j}. \quad (2)$$

We have used that in our approximating model, let’s call it the $Y$-model, any infected node in cluster $l$ infects any healthy node in cluster $j$ at rate equal to $Y_{w,l}^T Y_{h,j}/n$ (this is even true when $l = j$).

For this model, it is not so hard to determine a meta-stable state: we have to look for a vector $N_\infty \in \mathbb{R}^r_+$ such that for each component, the rate of increase by one is equal to the rate of decrease by one. In other words, using the previous rate equations, we get

$$\forall j \in \{1, \ldots, r\} : \frac{1}{n} \sum_{l=1}^{r} (n_j - N_j) N_l^\infty Y_{w,l}^T Y_{h,j} = N_j^\infty Y_{\delta,j}. \quad (3)$$

This is an $r$-dimensional vector equation, and we will consider the existence of a solution later. If $k < r$, we can also solve these equations in the following way: define the vector $V \in \mathbb{R}^k$ by

$$V = \frac{1}{n} \sum_{l=1}^{r} N_l^\infty Y_{w,l}. \quad (5)$$

Then (3) becomes

$$\forall j \in \{1, \ldots, r\} : (n_j - N_j^\infty) V^T Y_{h,j} = N_j^\infty Y_{\delta,j}. \quad (4)$$

This implies that

$$N_j^\infty = n_j \cdot \frac{V^T Y_{h,j}}{V^T Y_{h,j} + Y_{\delta,j}}. \quad (4)$$

All we need to do now is to determine the vector $V$. This can be done using the following equation:

$$V = \frac{1}{n} \sum_{l=1}^{r} n_l \cdot \frac{Y_{w,l} Y_{h,l}^T V}{Y_{h,l}^T V + Y_{\delta,l}}. \quad (5)$$

This is a $k$-dimensional vector equation, which can be solved efficiently numerically for relatively small $k$, if a solution exists.
3.2 Continuous approximation of the process \( N \)

Our Markov process \( N \), following the transfer rates given by (1) and (2), can in turn be approximated by a continuous Markov process in a classical way, by approximating a difference of Poisson processes by a Brownian motion with drift. For this it is relevant to center and rescale the process \( N \), so we define the deviations process \( D \) by

\[
D = \frac{N - N^\infty}{\sqrt{n}}.
\]

The reason for the \( \sqrt{n} \)-scaling will become clear later. The transfer rates for \( D \) are then given by

\[
D_j \rightarrow D_j + 1/\sqrt{n} \text{ at rate } \frac{1}{n} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)(n^{1/2}D_t + N_t^\infty)Y_{w,l}^{T}Y_{h,j}^{T} \tag{6}
\]

\[
D_j \rightarrow D_j - 1/\sqrt{n} \text{ at rate } (N_j^\infty + n^{1/2}D_j)Y_{\delta,j} \tag{7}
\]

Now suppose that \( n \) is large and we are at some time \( t \). Then there will be a lot of infection events and healing events in a relatively small time interval \([t, t + h]\). Also imagine that \( h \) is so small, that the transfer rates can be assumed to be constant during that time interval (this means that the changes in \( D \) during the time interval are negligible). Then we can define \( I_j(h) \) as the (random) number of nodes in cluster \( j \) that got infected during the time interval \([t, t + h]\), and \( H_j(h) \) as the number of nodes in cluster \( j \) that healed in that time interval. In good approximation these random variables will all be independent and Poisson distributed. We can explicitly calculate the expectation of these random variables, conditioned on \( D(t) \):

\[
\mathbb{E}(I_j(h) \mid D(t)) = h \frac{1}{n} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)(n^{1/2}D_t + N_t^\infty)Y_{w,l}^{T}Y_{h,j}^{T}
\]

and

\[
\mathbb{E}(H_j(h) \mid D(t)) = h(N_j^\infty + n^{1/2}D_j)Y_{\delta,j}.
\]

We can use this to calculate the expectation of \( dD(h) = D(t + h) - D(t) = n^{-1/2}(I(h) - H(h)) \):

\[
\mathbb{E}(dD_j(h) \mid D(t)) = h \left( \frac{1}{n^{3/2}} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)(n^{1/2}D_t + N_t^\infty)Y_{w,l}^{T}Y_{h,j}^{T} - (n^{-1/2}N_j^\infty + D_j)Y_{\delta,j} \right).
\]

Using Equation (3), this becomes

\[
\mathbb{E}(dD_j(h) \mid D(t)) = h \cdot \left( \frac{1}{n} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)D_t - D_jN_t^\infty \right)Y_{w,l}^{T}Y_{h,j}^{T} - D_jY_{\delta,j} \right).
\]

Conditionally, the components of \( dD(h) \) are independent, so we only have to determine the variances:

\[
\text{Var}(dD_j(h) \mid D(t)) = h \cdot \left( \frac{1}{n^{3/2}} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)(n^{1/2}D_t + N_t^\infty)Y_{w,l}^{T}Y_{h,j}^{T} + \frac{1}{n}(N_j^\infty + n^{1/2}D_j)Y_{\delta,j} \right)
\]

\[
= h \cdot \left( \frac{1}{n^{3/2}} \sum_{l=1}^{r} (n_j - N_j^\infty - n^{1/2}D_j)D_t - D_jN_t^\infty \right)Y_{w,l}^{T}Y_{h,j}^{T} + \frac{1}{n}(2N_j^\infty + n^{1/2}D_j)Y_{\delta,j} \right).
\]

The last equality follows from Equation (3). Note that we are adding the expectations (which equal the variances) of \( I_j(h) \) and \( H_j(h) \), which is why in this case, there is no cancelation of the main order terms, leading to the term \( 2N_j^\infty Y_{\delta,j}/n \).
It is useful to put these equations in matrix notation. Define \( n \) as the vector \((n_1, \ldots, n_r)\), and define the \( k \times r \) matrices

\[
Y_w = \left( Y_{w,1} \mid \cdots \mid Y_{w,r} \right) \quad \text{and} \quad Y_h = \left( Y_{h,1} \mid \cdots \mid Y_{h,r} \right).
\]

Then

\[
\mathbb{E}(dD(h) \mid D(t)) = h \cdot \left( \frac{1}{n} \operatorname{diag}(n - N^\infty - n^{1/2}D)Y_h^T Y_w D - \frac{1}{n} \operatorname{diag}(Y_h^T Y_w N^\infty) D - \operatorname{diag}(Y_\delta) D \right).
\]

Here we use the following notation for a vector \( a \in \mathbb{R}^p \):

\[
\operatorname{diag}(a) = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}.
\]

We introduce the \( r \times r \)-matrix \( M \):

\[
M = \frac{1}{n} \operatorname{diag}(n - N^\infty)Y_h^T Y_w - \frac{1}{n} \operatorname{diag}(Y_h^T Y_w N^\infty).
\]

Then we get

\[
\mathbb{E}(dD(h) \mid D(t)) = h \cdot \left( (M - \operatorname{diag}(Y_\delta)) D - \frac{1}{\sqrt{n}} \operatorname{diag}(D) Y_h^T Y_w D \right). \quad (8)
\]

For the conditional variance of \( dD(h) \) we get a similar formula:

\[
\operatorname{Var}(dD(h) \mid D(t)) = h \cdot \left( \frac{1}{\sqrt{n}} (M + \operatorname{diag}(Y_\delta)) D - \frac{1}{n} \operatorname{diag}(D) Y_h^T Y_w D + \frac{2}{n} \operatorname{diag}(Y_\delta) N^\infty \right). \quad (9)
\]

For large \( n \), it is natural to consider \( D \) as a continuous process. We have seen that \( dD(h) \) is distributed as the scaled difference of independent Poisson random variables, so it can be approximated by a normal distribution. Equations (8) and (9) then suggest to model \( D \) as the solution of the following stochastic vector differential equation:

\[
dD(t) = \left( (M - \operatorname{diag}(Y_\delta)) - \frac{1}{\sqrt{n}} \operatorname{diag}(D(t)) Y_h^T Y_w \right) D(t) dt + \operatorname{diag} \left( \sqrt{\frac{1}{\sqrt{n}} (M + \operatorname{diag}(Y_\delta)) D - \frac{1}{n} \operatorname{diag}(D) Y_h^T Y_w D + \frac{2}{n} \operatorname{diag}(Y_\delta) N^\infty} \right) dB(t). \quad (10)
\]

Here, \( B(t) \) is \( r \)-dimensional standard Brownian motion, and we use the notation

\[
\sqrt{a} = (\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_p})
\]

for \( a \in \mathbb{R}_+^p \).

Equation (10) can be used to efficiently simulate the time-evolution of \( D \), and in that way get information about the metastable distribution, but also about the relaxation time, for example. Since the SDE is non-linear, exact analysis is hard. However, for large \( n \) we can make some extra approximations.
3.3 Existence of the metastable solution

We can write Equation (3) in matrix form: define

\[ \tilde{A} = \frac{1}{n} \text{diag}(n/\delta) Y_h^T Y_w, \]

then Equation (3) becomes:

\[ 0 = \frac{1}{n} \text{diag}(n) Y_h^T Y_w N^\infty - \frac{1}{n} \text{diag}(N^\infty) Y_h^T Y_w N^\infty \]
\[ = \text{diag}(\delta)(\tilde{A} - I) N^\infty - \frac{1}{n} \text{diag}(N^\infty) Y_h^T Y_w N^\infty \]

We can prove that this has a positive solution when the positive \( r \times r \)-matrix \( \tilde{A} \) has a positive eigenvalue greater than 1. Now assume that all healing rates are the same: \( \Delta_i = \delta \). Then \( \tilde{A} \) becomes

\[ \tilde{A} = \frac{1}{\delta n} \text{diag}(n)^T Y_w. \]

Furthermore, suppose that each \( Z_i \) is very close to its cluster center, for example if \( r \) is close to \( n \). Define the map \( \text{Cl} : \{1, \ldots, n\} \to \{1, \ldots, r\} \), such that \( \text{Cl}(i) = j \) precisely when \( i \in C_j \). Then we see that

\[ A_{ii'}^T = H_i^T W_{i'} = \frac{1}{n} (Y_h^T Y_w)_{\text{Cl}(i)\text{Cl}(i')}. \]

So if \( \tilde{v} \in \mathbb{R}^r \) is the positive eigenvector of \( \tilde{A} \) with eigenvalue \( \lambda > 1 \), and we define \( v \in \mathbb{R}^n \) by

\[ v_i = \tilde{v}_{\text{Cl}(i)}, \]

then

\[ \frac{1}{\delta} (A^T v)_i = \frac{1}{\delta} \sum_{i'=1}^n A_{ii'}^T v_{i'} = \frac{1}{\delta} \sum_{j=1}^r \frac{n_{ij}}{n} (Y_h^T Y_w)_{\text{Cl}(i)j} \tilde{v}_j = (\tilde{A} \tilde{v})_{\text{Cl}(i)} = \lambda v_i. \]

We see that \( \lambda \) is also an eigenvalue of \( A^T/\delta \), which means that a metastable solution exists if the largest eigenvalue of \( A^T/\delta \) is larger than 1, exactly the same condition as NIMFA predicts in this case!

3.4 Linearising the SDE

Each node is characterised by its infectiousness, susceptibility and healing rate, captured in the vector \((W_i, H_i, \Delta_i)\). As \( n \) increases, the average infection rate between two nodes has to decrease, in order for a reasonable metastable state to exist (otherwise all nodes will be infected almost all of the time). We wish to study what happens for large \( n \), so we need to describe how the network grows with \( n \). Therefore we suppose that we can model the vectors \((n^{1/2}W_i, n^{1/2}H_i, \Delta_i)\) as a sample of size \( n \) from a distribution \( \mu \) on \( \mathbb{R}^{2k+1} \):

\[ \begin{pmatrix} n^{1/2}W_i \\ n^{1/2}H_i \\ \Delta_i \end{pmatrix} \sim \mu. \]

Note that the model only looks at the inner products of \( W_i \) and \( H_j \), so we can always make sure that \( W \) and \( H \) live at the same scale. Now we partition \( \mathbb{R}^{2k+1}_+ \) into \( r \) disjoint sets \( B_1, \ldots, B_r \) of positive
µ-mass, such that their “centers of mass” have small average distance to the other points in the set: we define
\[ \rho_j = \mu(B_j) \text{ and } \bar{Y}_j = \frac{1}{\rho_j} \int_{B_j} y\mu(dy). \]

Our clusters \(C_1, \ldots, C_r\) are now defined by
\[ i \in C_j \iff \left( \frac{n^{1/2}W_i}{\rho_j}, \frac{n^{1/2}H_i}{\rho_j}, \Delta_i \right) \in B_j. \]

Then \(n_j/n \approx \rho_j\),
\[ Y_{w,j} = \frac{1}{n_j} \sum_{i \in C_j} n^{1/2}W_i \approx \bar{Y}_{w,j} \text{ and } Y_{h,j} = \frac{1}{n_j} \sum_{i \in C_j} n^{1/2}H_i \approx \bar{Y}_{h,j} \]
and
\[ Y_{\delta,j} = \frac{1}{n_j} \sum_{i \in C_j} \Delta_i \approx \bar{Y}_{\delta,j}. \]

So the matrix \(Y^T_hY_w\) stabilises as \(n \to \infty\). Now look at Equation (3) and define \(\bar{N}^\infty \in [0, \rho_1] \times \ldots \times [0, \rho_r]\) by the equations
\[ \forall j \in \{1, \ldots, r\} : \sum_{l=1}^r (\rho_j - \bar{N}^\infty_j) \bar{N}^\infty_l Y^T_{w,l} \bar{Y}_{h,j} = \bar{N}^\infty \bar{Y}_{\delta,j}. \]

Dividing Equation (3) by \(n\) shows that for large \(n\),
\[ N^\infty \approx n\bar{N}^\infty. \]

When we take a closer look at the SDE (10), we can see that we only need to consider what happens to the matrix \(M\) as \(n \to \infty\):
\[ M = \frac{1}{n} \text{diag}(n - N^\infty)Y^T_hY_w - \frac{1}{n} \text{diag}(Y^T_hY_wN^\infty) \approx \text{diag}(\rho - \bar{N}^\infty)Y^T_hY_w - \text{diag}(Y^T_hY_w\bar{N}^\infty). \]

We can see that \(M\) also stabilises! We conclude which terms in the SDE (10) disappear as \(n \to \infty\), and which terms remain:
\[ dD(t) = (M - \text{diag}(Y_{\delta}))D(t)dt + \text{diag} \left( \sqrt{2\text{diag}(Y_{\delta})N^\infty/n} \right) dB(t). \]  

Note that the resulting SDE is a standard linear SDE, with a constant diagonal covariance factor! In fact, this SDE can be solved exactly: defining
\[ A = M - \text{diag}(Y_{\delta}) \text{ and } \Sigma = \text{diag} (2\text{diag}(Y_{\delta})N^\infty) / n, \]
we get
\[ D(t) = e^{tA}D(0) + \int_0^t e^{(t-s)A} \sqrt{\Sigma} dB(s). \]

Since we are interested in the metastable solution, we will have that \(A\) has eigenvalues with negative real part, so
\[ \lim_{t \to \infty} e^{tA}D(0) = 0. \]
Also, the asymptotic covariance will be given by

$$\Sigma_\infty = \lim_{t \to \infty} \text{Cov}(D(t)) = \int_0^\infty e^{\gamma t} \Sigma dt.$$  

(12)

There is another way to determine $\Sigma_\infty$: when we are in the metastable state, the distribution of $D(t)$ should not depend anymore on $t$. Now suppose that the metastable state is given by

$$D(t) \sim N(\mu, \Sigma_\infty).$$

Use (11) to see that $D(t + dt) = D(t) + dD(t)$ also has a normal distribution. Furthermore,

$$\mathbb{E}(dD(t)) = A\mu dt + \text{diag} \left( \sqrt{2\text{diag}(Y_\delta)N_\infty/n} \right) \mathbb{E}(dB(t)) = A\mu dt.$$  

So $\mu = 0$ implies that the expectation is stable. Furthermore, using that $dB(t)$ is independent of the past, we get

$$\text{Cov}(D(t + dt)) = \text{Cov}(D(t) + AD(t)dt) + \text{diag} \left( \sqrt{2\text{diag}(Y_\delta)N_\infty/n} \right) \text{Cov}(dB(t)) \text{diag} \left( \sqrt{2\text{diag}(Y_\delta)N_\infty/n} \right)$$

$$= (I + Adt)\Sigma_\infty (I + A^T dt) + \Sigma dt$$

$$= \Sigma_\infty + (A\Sigma_\infty + \Sigma_\infty A^T + \Sigma)dt + O(dt^2).$$

So metastability implies that

$$A\Sigma_\infty + \Sigma_\infty A^T + \Sigma = 0.$$  

It is not so hard to see that $\Sigma_\infty$ defined by (12) satisfies this matrix equation.

We conclude that the metastable distribution $N = N_\infty + \sqrt{n}D$ will be an $r$-dimensional normal distribution, with expectation equal to $N_\infty$ (since $D$ has expectation 0), and covariance matrix $n\Sigma_\infty$.

We can also look at the correlation between two nodes $i_0$ and $i_1$. Define $X_i = 1$ if node $i$ is infected, and otherwise $X_i = 0$. Now suppose node $i_0$ is an element of cluster $j_0$, and node $i_1$ is an element of cluster $j_1$. If $j_0 = j_1$, we still assume that $i_0 \neq i_1$. We know that

$$N_j = \sum_{i \in C_j} X_i = N_{j_0} + \sqrt{n} \cdot D_j.$$  

Also, all nodes in cluster $j$ are approximately indistinguishable, so

$$\mathbb{E}(X_{i_0}) = N_{j_0}^\infty / n_{j_0} \approx \rho_{j_0}.$$  

The indistinguishability also implies that

$$\text{Cov} \left( \sum_{i \in C_{j_0}} X_i, \sum_{i' \in C_{j_1}} X_{i'} \right) = n_{j_0}n_{j_1} \text{Cov}(X_{i_0}, X_{i_1}) + 1_{\{j_0 = j_1\}} n_{j_0} \text{Cov}(X_{i_0}, X_{i_0} - X_{i_1}) = n\Sigma_{j_0,j_1}.$$  

We conclude that for $j_0 \neq j_1$,

$$\text{Corr}(X_{i_0}, X_{i_1}) = \frac{n\Sigma_{j_0,j_1}}{n_{j_0}n_{j_1} \sqrt{\rho_{j_0}(1 - \rho_{j_0})\rho_{j_1}(1 - \rho_{j_1})}}$$  

and when $j_0 = j_1$ we get

$$\text{Corr}(X_{i_0}, X_{i_1}) = \frac{n\Sigma_{j_0,j_0} - n_{j_0}\rho_{j_0}(1 - \rho_{j_0})}{n_{j_0}^2 \rho_{j_0}(1 - \rho_{j_0})}.$$
Summarising, when $N^\infty$ is of order $n$, then the fluctuations are normally distributed and of order $\sqrt{n}$. Also, the correlations between different nodes can be calculated exactly, and they are of order $1/n$. These individual correlations decay, but they cannot be neglected when calculating correlations between the number of infected nodes in two clusters, since this is determined by the covariance matrix $\Sigma^\infty$, which in general is not diagonal.

3.5 Choosing the number of clusters $r$

When applying our method to a particular example, we face two problems: we have to choose the dimension $k$ of our non-negative matrix factorization of $A$, and we have to choose the number of clusters $r$. The first problem is hard, and it depends on how well $A$ can be approximated by the product of two low rank matrices. Choosing the number of clusters is less delicate: if we choose a lot of clusters, but a subset of the cluster centers are very close together, then the clusters themselves become indistinguishable: this means that the total number of infected nodes in the union of the respective subset of clusters behaves as though it is one cluster, also in the approximating SDE. This guarantees a certain stability: choosing more clusters does not really change the predicted behavior of a fixed subset of nodes. We therefore recommend choosing a lot of clusters (maybe even $r = n$), since the bias effect is less severe, whereas the effect on the variance is limited, especially when looking at larger collections of nodes. Clearly, on a micro-level (so when looking at a few nodes), the approximation will not get much better.

References

[1] E. Cator, P. Van Mieghem, *Nodal infection in Markovian susceptible-infected-susceptible and susceptible-infected-removed epidemics on networks are non-negatively correlated*, Physical Review E 89 (5), 052802, 2014.