A linear algebraic approach to understanding the basic reproduction number

Andrew F. Brouwer

An understanding of linear algebra is useful for infectious disease modeling. Many analyses—the basic reproduction number and initial growth analysis, among others—are conveniently calculated using linear algebra techniques. However, knowing how to invert and multiply matrices as one might in a basic linear algebra class does not necessarily translate into an understanding of what the quantities mean, and having a heuristic understanding of the epidemiologic concepts is not always enough to be confident in ones mathematical techniques. Here, we give geometric and practical interpretations of the next generation method for calculating the basic reproduction number. I use the notation of van den Driessche and Watmough.

This essay is aimed at people who have seen the epidemiologic concept of the basic reproduction number and some linear algebra but would like a more explicit connection between the two.

The Basic Reproduction Number

The *basic reproduction number* R_0 is an important quantity in infectious disease systems epidemiology, defined as the average number of secondary cases arising from an typical primary case in an entirely susceptible population. The basic reproduction number acts as a threshold value that controls the local stability of the disease-free equilibrium: if $R_0 < 1$, the disease will die off quickly, while if $R_0 > 1$, the disease will become epidemic. In practice, R_0 is calculated as a threshold parameter that may not precisely correspond to the number of secondary cases per infection, especially in the case of an environmentally transmitted infections. The values of R_0 vary greatly by disease, ranging from close to 1 for seasonal influenza to 5–7 for smallpox and polio to 12–18 for measles and pertussis, but are also dependent on some attributes of the population. Mathematical modeling is often used to estimate the basic reproductive number; in fact, the basic reproduction number is widely considered the most useful contribution of mathematics to epidemiology.

In the basic SIR system,

$$\begin{aligned}\dot{S} &= -\beta S I,\\ \dot{I} &= \beta S I - \gamma I,\\ \dot{R} &= \gamma I,\end{aligned} \tag{1}$$

the basic reproduction number is given by the ratio of the transmission rate to the recovery rate β/γ. This is the rate of an infectious person's contacts with susceptible people times the average time infected; this naturally gives the average number of people infected by that first person. However, more complicated systems may not have such an immediate way to understand R_0.

Several methods exist for calculating R_0 in more complex systems. One of the most rigorous and commonly used approaches is the next generation method. Denote the vector of states by x and the disease free equilibrium by x_0. For each infected compartment i, let $\mathcal{F}_i(x)$ be the rate at which previously uninfected people enter compartment i. Let $\mathcal{V}_i(x)$ be the rate of transfer of individuals out of compartment i minus the rate of transfer into compartment i. Then

$$\frac{dx_i}{dt} = \mathcal{F}_i(x) - \mathcal{V}_i(x). \tag{2}$$

Let F and V by the Jacobian (generalized derivative of multivalued, mulivariate functions) matirx of $\mathcal{F}$ and $\mathcal{V}$, i.e. the matrices whose entries are

$$F_{ij} = \frac{\partial \mathcal{F}_i(x)}{\partial_{x_j}}\Big|_{x=x_0}, \tag{3}$$

$$V_{ij} = \frac{\partial \mathcal{V}_i(x)}{\partial_{x_j}}\Big|_{x=x_0}. \tag{4}$$

The matrix $K = FV^{-1}$ is called the next generation matrix. Why this form, and how does one interpret the entries of this matrix? From van den Driessche and Watmough (2002):

> To interpret the entries of FV^{-1} and develop a meaningful definition of R_0, consider the fate of an infected individual introduced into compartment k of a disease free population. The (j, k) entry of V^{-1} is the average length of time this individual spends in compartment j during its lifetime, assuming that the population remains near the DFE and barring reinfection. The (i, j) entry of F is the rate at which infected individuals in compartment j produce new infections in compartment i. Hence, the (i, k) entry of the product FV^{-1} is the expected number of new infections in compartment i produced by the infected individual originally introduced into compartment k.

Then, R_0 is defined to be the spectral radius—that is, the largest eigenvalue—of the matrix $K = FV^{-1}$, often denoted $\rho(K)$.

Why is it that this largest eigenvalue controls the behavior of the infectious disease system? Let us consider the geometric interpretation of the the eigenvalues and eigenvectors of a matrix K. A circle (or n-sphere) is represented by the set of all vectors with norm 1, that is $\{x : ||x|| = 1\}$. Apply matrix K to this set of vectors, we get an (n-dimensional) ellipse if K is non-singular. This ellipse has a major and minor axis corresponding the the eigenvectors of matrix $(K^{-1})'K^{-1}$. (Note: for any matrix M, the square root of the eigenvalues of $M'M$ are known as the singular values of M). The stretch along each of these axes is the reciprocal of the square root of the corresponding eigenvalue of $(K^{-1})'K^{-1}$, i.e. the corresponding singular value of K. The eigenvectors of K, on the other hand, represent those vectors that did not change angle when K was applied to the circle. Since one of the eigenvectors of K stretches by the largest eigenvalue—$\rho(K)$—then the greatest magnitude of stretch of any vector—called the operator norm $||K||$—can be no smaller than $\rho(K)$. That is,

$$\max_k |\lambda_k| =: \rho(K) \leq ||K|| := \max_{||x||=1} ||Kx||. \tag{5}$$

Now, what happens when we apply matrix K to the initial circle multiple times? The ellipse becomes more exaggerated. But, the relative contributions of the eigenvectors of K changes. This tells us about the long-term behavior of the system. But, we are interested in the average behavior. That is, we need to scale our ellipse. We are interested in $\sqrt[m]{||K^m||}$ as m increases. As we do this scaling, the points converge along the eigenvector corresponding to $\rho(K)$. Gelfand's Formula states that

$$\rho(K) = \lim_{m\to\infty} \sqrt[m]{||K^m||}. \tag{6}$$

Hence the spectral radius gives the average long-run behavior. The influence of the other eigenvalues dies out quickly, so in the long-run average, the behavior is the controlled by only the largest eigenvalue.

In the infectious disease context, each application of K is another generation of infected people. Although a specific outbreak will be influenced by the initial conditions, the average behavior, i.e. the expected number new infected people given a single infectious, is controlled only by the spectral radius.

Examples

Computational example

Consider the matrix

$$A = \begin{bmatrix} 1.58 & 0.84 \\ 0.14 & 1.72 \end{bmatrix}.$$

If this is a next generation matrix, one individual of type 1 will make on average 1.58 infections of type 1 and 0.14 infection of type 2. We can see that different intial conditions will produce different numbers of infected people in the second generation. Starting with one person of type 1 will result in 1.72 infections on average in the next generation, but starting with one person of type 2 will result in 2.56 infections.

Matrix A has eigenvalues 2 and 1.3 and eigenvectors $(2, 1)$ and $(3, -1)$. Thus, we have the decomposition

$$A = \begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1.3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}^{-1}. \tag{7}$$

How do we describe the ellipse $\{Ax : |x| = 1\}$? Let $z = Ax$. Then,

$$\begin{aligned} z &= Ax, \\ A^{-1}z &= x, \end{aligned} \tag{8}$$

and

$$\begin{aligned} 1 &= x'x, \\ &= (A^{-1}z)'(A^{-1}z), \\ &= z'(A^{-1})'A^{-1}z. \end{aligned} \tag{9}$$

We have the quadratic form $z'Qz$ where $Q = (A^{-1})'A^{-1}$. Since Q is symmetric, it has an orthogonal (perpendicular) basis of eigenvectors, which form the major and minor axes of the ellipse. The lengths of these axes are the reciprocals of the square roots of its eigenvalues. The operator norm of A is the reciprocal of the square root of the smallest eigenvalue of Q (the same as the square root of the largest eigenvalue of $A'A$, the largest singluar value). In this example, we have $\rho(A) = 2$ but $||A|| = 2.18$.

In Figure 1a, we plot $\{Ax : |x| = 1\}$ with the major and minor axes and the eigenvectors of A. The eigenvectors do not correspond to the major and minor axes. The axes are orthogonal. In Figure 1b, we plot $\{A^m x : |x| = 1\}$ for $0 \leq m \leq 6$. The ellipse becomes more exaggerated, and appears to rotate slightly. In Figure 1c, we plot $\{\frac{A^m x}{||A^m x||} \cdot \sqrt[m]{||A^m x||} : |x| = 1\}$. The norm of these transformations is $\sqrt[m]{||A^m||}$, and we see that the direction and magnitude of the largest stretch of this "average behavior" function, moves toward the eigenvector and dominant eigenvalue of the original transformation. Since the greatest stretch lies outside of the the circle of radius 1, there will be an outbreak.

We repeat this exercise with two more matrices. You may like to write their eigenvalue decomposition yourself.

$$B = \begin{bmatrix} 1.30 & 1.20 \\ 0.15 & 1.00 \end{bmatrix} \tag{10}$$

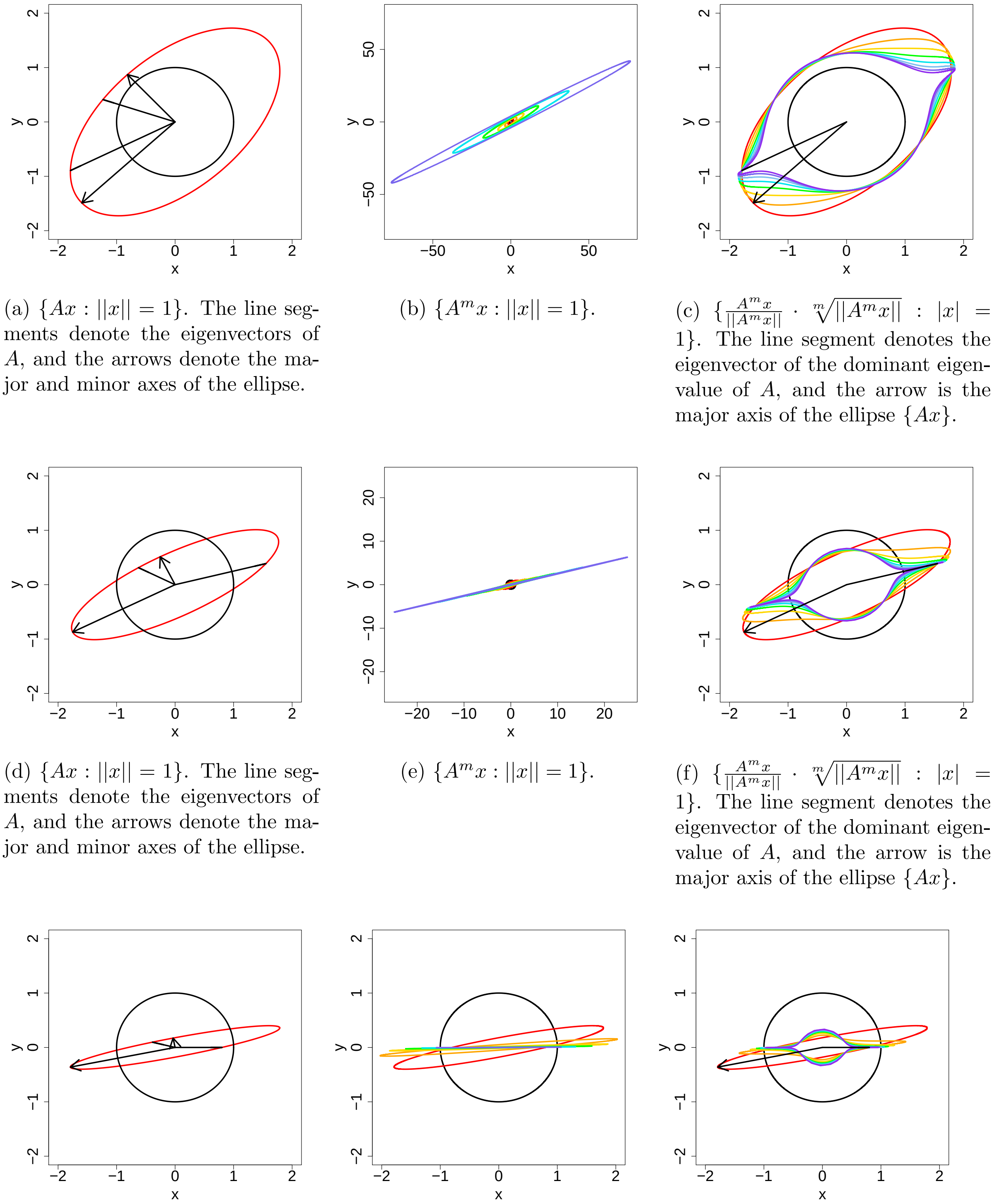
$$C = \begin{bmatrix} 0.80 & 1.60 \\ 0 & 0.40 \end{bmatrix} \tag{11}$$

Matrix B has eigenvalues 1.6 and 0.7, and we expect an outbreak. Matrix C, has eigenvalues 0.8 and 0.4, so, despite the fact that infections of type 2 produce an average of 2 infections in the next generation, we do not expect an outbreak. Figures 1 d, e, f and Figure 1 g, h, i are analagous to Figures 1 a, b, and c for matrices B and C.

SEIR model with demography

One simple extension of the SIR model is the SEIR model, in which we add an exposed-but-not-yet-infectious compartment E. The equations—with demography—are

$$\begin{aligned} \dot{S} &= \mu - \beta SI - \mu S, \\ \dot{E} &= \beta SI - \sigma E - \mu E, \\ \dot{I} &= \sigma E - \gamma I - \mu I, \\ \dot{R} &= \gamma I - \mu R. \end{aligned} \tag{12}$$



(a) $\{Ax : ||x|| = 1\}$. The line segments denote the eigenvectors of A, and the arrows denote the major and minor axes of the ellipse.

(b) $\{A^m x : ||x|| = 1\}$.

(c) $\{\frac{A^m x}{||A^m x||} \cdot \sqrt[m]{||A^m x||} : |x| = 1\}$. The line segment denotes the eigenvector of the dominant eigenvalue of A, and the arrow is the major axis of the ellipse $\{Ax\}$.

(d) $\{Ax : ||x|| = 1\}$. The line segments denote the eigenvectors of A, and the arrows denote the major and minor axes of the ellipse.

(e) $\{A^m x : ||x|| = 1\}$.

(f) $\{\frac{A^m x}{||A^m x||} \cdot \sqrt[m]{||A^m x||} : |x| = 1\}$. The line segment denotes the eigenvector of the dominant eigenvalue of A, and the arrow is the major axis of the ellipse $\{Ax\}$.

(g) $\{Ax : ||x|| = 1\}$. The line segments denote the eigenvectors of A, and the arrows denote the major and minor axes of the ellipse.

(h) $\{A^m x : ||x|| = 1\}$.

(i) $\{\frac{A^m x}{||A^m x||} \cdot \sqrt[m]{||A^m x||} : |x| = 1\}$. The line segment denotes the eigenvector of the dominant eigenvalue of A, and the arrow is the major axis of the ellipse $\{Ax\}$.

Figure 1

Here, μ is the birth/death rate, β is the contact rate, $1/\sigma$ is average time in the exposed compartment, and $1/\gamma$ is the average time spent in the infectious compartment. Then there are two infected compartments, I and E. We write

$$\mathcal{F} = \begin{bmatrix} \beta SI \\ 0 \end{bmatrix} \tag{13}$$

$$\mathcal{V} = \begin{bmatrix} (\sigma+\mu)E \\ (\gamma+\mu)I - \sigma E \end{bmatrix} \tag{14}$$

and

$$F = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} \tag{15}$$

$$V = \begin{bmatrix} \sigma+\mu & 0 \\ -\sigma & \gamma+\mu \end{bmatrix} \tag{16}$$

Note that the ith component of the diagonal of V is rate of leaving compartment i (the sum of exponential rates is gives the rate of the first event, whichever that ends up being). The other elements in the corresponding column are the rates of movement to the other infected compartments. The sum of a column gives the rate of becoming not infected (whether through recovery or death). Then,

$$V^{-1} = \begin{bmatrix} \frac{1}{\sigma+\mu} & 0 \\ \frac{\sigma}{(\sigma+\mu)(\gamma+\mu)} & \frac{1}{\gamma+\mu} \end{bmatrix} \tag{17}$$

$$FV^{-1} = \begin{bmatrix} \frac{\beta\sigma}{(\sigma+\mu)(\gamma+\mu)} & \frac{\beta}{\gamma+\mu} \\ 0 & 0 \end{bmatrix} \tag{18}$$

Why is the second row of the next generation matrix all zeros? Well, this row corresponds to the number of new people in compartment I given a person in either E or I *near the disease-free equilibrium.* Any new infections produce exposed people, not infected people directly. Since we are considering only a neighborhood of the disease-free equilibrium, we do not consider events beyond the initial infection.

Now, the eigenvalues of the next generation matrix are 0 and

$$\rho(FV^{-1}) = \left(\frac{\sigma}{\sigma+\mu}\right)\left(\frac{\beta}{\gamma+\mu}\right). \tag{19}$$

We see that the basic reproduction number of the SEIR model is the basic reproduction number of the SIR model—$\beta/(\gamma+\mu)$—times the fraction of exposed who go on to being infectious—$\sigma/(\sigma+\mu)$. We also see that as $\sigma \to \infty$, that is, as transition out of the exposed compartment becomes instantaneous, the basic reproduction number converges to that of the SIR model, as expected.

A treatment-compliance model

For the purposes of illustrating some of the complexities of interpreting V^{-1}, we also consider a two compartment model. This model has two classes of infected people, and individuals can go back and forth between compartments. We can think of this as a simplified model of compliance of AIDS treatment. Here, treatment might reduce infectivity and reduce the mortality rate. To reduce complexity for this example, we ignore demography. Here β_1 and β_2 are contact rates times transmission probabilities, ϕ is the rate of going on treatment, π is the rate of treatment lapse, and ν_1 and ν_2 are the disease-related death rates.

$$\begin{aligned}\dot{S} &= -S(\beta_1 I + \beta_2 T)\\ \dot{I} &= S(\beta_1 I + \beta_2 T) - \phi I + \pi T - \nu_1 I\\ \dot{T} &= \phi I - \pi T - \nu_2 T\end{aligned} \tag{20}$$

We have

$$\mathcal{F} = \begin{bmatrix} S(\beta_1 I + \beta_2 T)\\ 0\end{bmatrix}, \tag{21}$$

$$\mathcal{V} = \begin{bmatrix} (\nu_1 + \phi)I - \pi T\\ (\nu_2 + \pi)T - \phi I\end{bmatrix}, \tag{22}$$

and

$$F = \begin{bmatrix} \beta_1 & \beta_2\\ 0 & 0\end{bmatrix}, \tag{23}$$

$$V = \begin{bmatrix} \nu_1 + \phi & -\pi\\ -\phi & \nu_2 + \pi\end{bmatrix}. \tag{24}$$

Then,

$$V^{-1} = \begin{bmatrix} \frac{\nu_2+\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi}\\ \frac{\phi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{\nu_1+\phi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi}\end{bmatrix}. \tag{25}$$

Here, V^{-1} is much harder to interpret than in the previous example. The determinant of V, namely $(\nu_1+\phi)(\nu_2+\pi)-\phi\pi$, is more complicated and does not cancel with terms in the numerator. Indeed, how are we to interpret the determinant?

Consider an individual with an untreated infection. They can either die with probability $\frac{\nu_1}{\nu_1+\phi}$ or they can start treatment with probability $\frac{\phi}{\nu_1+\phi}$ (incidentally, thinking in these terms—identifting jumping probabilities—is the first step to transitioning to a stochastic framework). Similarly, a person on treatment can die with probability $\frac{\nu_2}{\nu_2+\pi}$ or they can stop taking their treatment with probability $\frac{\pi}{\nu_2+\pi}$. The probability of starting in one compartment, jumping to the other, and then jumping back is

$$x := \frac{\phi\pi}{(\nu_1+\phi)(\nu_2+\pi)}.$$

The probability of jumping back and forth twice is x^2, and so on. Given that one starts out not on treatment, what is the expected number of times that one will not be on treatment? Let us count the times. We start with one visit since one starts in that compartment. Then, we must add one for each additional visit, times the probability that the visit occurs. So, the expected number of visits is

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}.$$

We can arrive at this number using some basic probability theory as well. Let X, a random number, be the number of visits to compartment I, and let Z be the event of a return visit to compartment I. Using the law of total expectation

$$\begin{aligned} E[X] &= 1 + P(Z)E[X|Z] + (1-P(Z))E[X|\bar{Z}] \\ &= 1 + xE[X] + 0 \end{aligned} \tag{26}$$

Solving for $E[X]$, we get $E[X] = \frac{1}{1-x}$.

In terms of our model, this expected number of visits is

$$\frac{1}{1-\frac{\phi\pi}{(\nu_1+\phi)(\nu_2+\pi)}} = \frac{(\nu_1+\phi)(\nu_2+\pi)}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \tag{27}$$

How much time does one spend in the untreated compartment? Well, we expect there to be $\frac{(\nu_1+\phi)(\nu_2+\pi)}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi}$ visits, each lasting $\frac{1}{\nu_1+\phi}$. Thus, if one starts in the untreated compartment, one expects to spend

$$\frac{(\nu_1+\phi)(\nu_2+\pi)}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \cdot \frac{1}{\nu_1+\phi} = \frac{\nu_2+\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \tag{28}$$

much time there over the course of one's lifetime. This, of course, is $V_{1,1}$.

There is a nice graph-theoretic way of approaching this interpretation as well. Let A be the matrix whose entries $a_{i,j}$ are the probabilies of moving from compartment j to compartment i; this is the adjacency matrix of the directed graph of the compartments, weighted by transition probability (or, depending on your definition of the adjacency matrix, its transpose). Then

$$A = \begin{bmatrix} 0 & \frac{\pi}{\nu_2+\pi} \\ \frac{\phi}{\nu_1+\phi} & 0 \end{bmatrix}. \tag{29}$$

Then,

$$I + A + A^2 + A^3 + \cdots = (I-A)^{-1} \tag{30}$$

is a matrix whose entries m_{ij} give the expected number of visits to compartment i if one starts in compartment j. Thus, we can write V^{-1} as the product of waiting times and expected number of visits.

$$\begin{aligned} V^{-1} &= \begin{bmatrix} \frac{1}{\nu_1+\phi} & 0 \\ 0 & \frac{1}{\nu_2+\pi} \end{bmatrix} (I-A)^{-1} \\ &= \begin{bmatrix} \frac{1}{\nu_1+\phi} & 0 \\ 0 & \frac{1}{\nu_2+\pi} \end{bmatrix} \begin{bmatrix} \frac{(\nu_1+\phi)(\nu_2+\pi)}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{(\nu_1+\phi)\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \\ \frac{\phi(\nu_2+\pi)}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{(\nu_1+\phi)(\nu_2+\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\nu_2+\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{\pi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \\ \frac{\phi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} & \frac{\nu_1+\phi}{(\nu_1+\phi)(\nu_2+\pi)-\phi\pi} \end{bmatrix}. \end{aligned} \tag{31}$$

Type and target reproduction numbers

This section discusses some more advanced extensions of the basic reproduction number, namely the type and target reproduction numbers. The basic reproduction number has implications for infection control: if a fraction of the population greater than $1-\frac{1}{R_0}$ is permanently protected from infection (e.g. through immunization), the infection cannot become epidemic. The concept of the basic reproductive number, at least in these infection control terms, can be extended to examine infection control in populations with multiple subgroups using the type and target reproductive numbers, which provide R_0-like threshold quantities under the assumption that only a specific population group or transmission pathway is being controlled. Suppose that there are n host types. Let $K = FV^{-1}$ be the next generation matrix, and P_i the projection matrix with $P_{ii} = 1$ and all other entries 0. Then, if $\rho((I-P_i)K) < 0$, that is, if the other host groups are not self-sustaining disease reservoirs, the infection can be controlled by protecting a greater fraction than $1-\frac{1}{T_i}$ of host type i, where

$$T_i = e_i' K(I-(I-P_i)K)^{-1} e_i \tag{32}$$

is called the *type reproduction number* for host type i. It is known that $T_i > 1$ if and only if $R_0 > 1$ (Roberts and Heesterbeek, 2003). If $\rho((I-P_i)K) \geq 1$, then another host type acts as a reservoir for infection, the infection cannot be controlled only through intervention on host type i, and T_i is not defined. If several host types are to be controlled simultaneously, the type reproduction number is defined as

$$M_\ell = E_\ell' K(I-(I-P_\ell)K)^{-1} E_\ell, \tag{33}$$

where $(E_\ell)_{ii} = (P_\ell)_{ii} = 1$ for $i \in \ell$ and 0 otherwise. The type reproduction number is especially of interest for vector-borne and other multiple-species infections. The type reproduction number may be further generalized to consider individual entries of the next generation matrix, not just whole rows. For any matrix A such that $A_{ij} = K_{ij}$ or 0, then the *target reproduction number* for the nonzero entries of A is

$$T_A = \rho(A((I-(K-A))^{-1}). \tag{34}$$

The target reproduction number has the same properties as the type reproduction number (Shuai et al., 2013).

Selected references

On the next generation method for calculating the basic reproduction number:

- P. van den Driessch, J. Watmough. 2002. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences.* 180: 29–48

- P. van den Driessch, J. Watmough. 2008. Further notes on the basic reproduction number. *Mathematical Epidemiology.* Springer Berlin Heidelberg.

- O. Diekmann, J. A. P. Heesterbeek, M. G. Roberts. 2010. The construction of next-generation matrices for compartmental epidemic models. Journal of the Royal Society, Interface. 7(47): 873–885.

A graph-theoretic alternative to the next generation matrix:

- T. de-Camino-Beck, M. A. Lewis, P. van den Driessche. 2009. A graph-theoretic method for the basic reproduction number in continuous time epidemiological models.. *Journal of Mathematical Biology.* 59(4): 503–16.

On the type and target reproduction numbers:

- M. G. Roberts, J. A. P. Heesterbeek. 2003. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society, B.* 270(1522): 1359–64.

- J. A. P. Heesterbeek, M. G. Roberts. 2007. The type-reproduction number T in models for infectious disease control. *Mathematical Biosciences.* 206(1): 3–10.

- Z. Shuai, J. A. P. Heesterbeek, P. van den Driessch. 2013. Extending the type reproduction number to infectious disease control targeting contacts between types. *Journal of Mathematical Biology.* 67(5): 1067–82.