1999

Computing approximate stationary distributions for discrete Markov processes with banded infinitesimal generators

Borges, Carlos F.; Peters, Craig S.

Journal of Applied Probability, Vol. 36, pp. 1086-1100, 1999.
https://hdl.handle.net/10945/42183

J. Appl. Prob. **36,** 1086–1100 (1999)
Printed in Israel

COMPUTING APPROXIMATE STATIONARY DISTRIBUTIONS FOR DISCRETE MARKOV PROCESSES WITH BANDED INFINITESIMAL GENERATORS

CARLOS F. BORGES,* *Naval Postgraduate School, Monterey*
CRAIG S. PETERS,** *General Electric Co., New York*

Abstract

We develop an algorithm for computing approximations to the stationary distribution of a discrete birth-and-death process, provided that the infinitesimal generator is a banded matrix. We begin by computing stationary distributions for processes whose infinitesimal generators are Hessenberg. Our derivation in this special case is different from the classical case but it leads to the same result. We then show how to extend these ideas to processes where the infinitesimal generator is banded (or half-banded) and to quasi-birth–death processes. Finally, we give an example of the application of this method to a nearly completely decomposable Markov chain to demonstrate the general applicability of the technique.

Keywords: Homogeneous complement; singular value decomposition

AMS 1991 Subject Classification: Primary 60J
Secondary 65F

1. Introduction

A birth and death process is characterized by a population of individuals whose number changes according to the outcome of two other stochastic processes, consisting of births which increase the population and deaths which decrease the population. The transition probabilities for these two processes can, in general, depend on both time and population size. The model for this stochastic dynamical system is usually described by a master equation for the transition probability for population size at a given time.

Let $N(t)$ denote the population size at time t and define the transition probability for $N(t)$ as $P(n, t) = \Pr\{N(t) = n \mid N(t') = n'\}$. The transition probabilities for births and deaths are usually modeled as Markovian processes by assuming that

$$r(n, t) = \Pr\{N(t + dt) = k + 1 \mid N(t) = k\},$$
$$l(n, t) = \Pr\{N(t + dt) = k - 1 \mid N(t) = k\},$$

where dt represents the fundamental time unit. Time can be treated as either a discrete or continuous variable. In many situations it is reasonable and convenient to model these transition probabilities as being time independent and to assume that the probability that more

Received 29 June 1995; revision received 29 November 1996.
* Postal address: Code MA/BC, Naval Postgraduate School, Monterey, CA, 93943, USA.
Email address: borges@nps.navy.mil
** Postal address: General Electric Co., Building K1, SC1D, One Research Circle, Niskayina, NY 12309, USA.

than a single birth or death occurs in the fundamental time unit is zero. With these assumptions the master equation for $P(n, t)$ can be written as

$$P(n, t + dt) = r(n - 1)P(n - 1, t) + l(n + 1)P(n + 1, t) + [1 - (r(n) + l(n))]P(n, t). \tag{1}$$

The first approach to investigating (1) is to assume that

$$\lim_{t \to \infty} P(n, t + dt) - P(n, t) = 0,$$

and write the equation for the stationary transition probabilities, $p(n)$, the probability that the population will eventually stabilize at n individuals, as

$$0 = -(r(n) + l(n))p(n) + r(n - 1)p(n - 1) + l(n + 1)p(n + 1). \tag{2}$$

This *stochastic balance equation* leads to the *infinitesimal generator* for the discrete-time Markov process, which has the following stationary distribution

$$p(n) = p(0) \prod_{j=1}^{n} \frac{r(j - 1)}{l(j)}. \tag{3}$$

In this paper we develop an algorithm for computing the solution to this problem and show how it can be generalized to compute approximate solutions for birth-and-death processes where both multiple births and multiple deaths can occur in the fundamental time unit. These methods are developed by converting stochastic balance equations like (2) to matrix form and applying techniques from linear algebra.

The virtue of the linear algebra approach is two-fold. First, the use of finite precision calculations in linear algebra is well understood and many algorithms have been developed that can control the ill effects of round-off and other errors. Second, generalizations of the simple birth and death model give rise to generalizations of the master equations (1) and (2) for which solutions are not known. In this case, the methods we develop are still applicable. To motivate what follows, we now show how (3) results from solving an appropriate linear system.

2. A matrix formulation

We are interested in finding the stationary distribution $p(n)$ which satisfies (2). We begin by converting (2) to matrix form. In what follows, vectors will be denoted by lower-case bold letters and will be assumed to be column vectors. We shall denote by $\mathbf{0}$ the vector of all zeros, by $\boldsymbol{e}$ the vector of all ones, and by $\boldsymbol{e}_i$ the ith axis vector, a vector whose ith element is 1 and all others are zero. The sizes of these vectors, when they appear, shall be taken from context.

To proceed, we represent the stationary distribution $p(n)$ as a semi-infinite column vector $\boldsymbol{p}$ whose ith element $p_i = p(i - 1)$ where $i = 1, 2, \ldots, \infty$. Note that if there were a maximum population size N then $\boldsymbol{p}$ would be an element of $\mathbb{R}^{N+1}$.

The infinitesimal generator matrix Q^T is a semi-infinite tridiagonal matrix, with entries

$$Q^T_{i,j} = \begin{cases} -(r(i - 1) + l(i - 1)) & \text{if } i = j \\ r(j - 1) & \text{if } i - 1 = j \\ l(j - 1) & \text{if } i + 1 = j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

With these definitions, equation (2) becomes

$$Q^T \boldsymbol{p} = \mathbf{0}. \tag{5}$$

And we see that Q^T must be rank-deficient and $\boldsymbol{p}$ in its null-space (or equivalently, $\boldsymbol{p}$ is an eigenvector of Q^T associated with the eigenvalue $\lambda = 0$). Equation (5) implies that any element of the null-space of Q^T solves the equation. To get the stationary distribution we normalize using the law of total probability, so that $\boldsymbol{e}^T \boldsymbol{p} = 1$.

3. Truncated solutions

It is not generally possible to solve semi-infinite systems of equations, or find eigenvectors of semi-infinite matrices, so we consider truncating the system of equations. Graphically, we truncate the semi-infinite system by partitioning it in the following way:

$$\left[\begin{array}{ccccc|ccc} \times & \times & 0 & & & & & \\ \times & \ddots & \ddots & \ddots & & & & \\ 0 & \ddots & & \times & & & & \\ & \ddots & \times & \times & \times & & & \\ \hline & & & \times & \times & \times & & \\ & & & & \times & \ddots & \ddots & \\ & & & & & \ddots & & \end{array}\right] \left[\begin{array}{c} p(0) \\ p(1) \\ \vdots \\ p(n-1) \\ \hline p(n) \\ \vdots \end{array}\right] = \mathbf{0},$$

and then dropping all but the first n equations.

Letting Q_n^T be the principal $n \times n$ sub-matrix of Q^T, and $\boldsymbol{p}_n$ be the principal n-element sub-vector of $\boldsymbol{p}$, we get

$$Q_n^T \boldsymbol{p}_n = -l(n)p(n)\boldsymbol{e}_n.$$

In effect, this is a matrix representation of the first n equations from (5). Letting $\boldsymbol{p}_n = l(n)p(n)\boldsymbol{f}_n$ and rearranging yields

$$Q_n^T \boldsymbol{f}_n = -\boldsymbol{e}_n. \tag{6}$$

So, provided that Q_n^T is non-singular and that $l(n)p(n) \neq 0$, we can solve directly for a scalar multiple of the truncated stationary distribution $\boldsymbol{p}_n$.

In particular, the truncated infinitesimal generator matrix is

$$Q_n^T = \begin{bmatrix} -r(0) & l(1) & & & & \\ r(0) & -r(1) - l(1) & l(2) & & & \\ & r(1) & & \ddots & & \\ & & \ddots & & l(n-1) & \\ & & & r(n-2) & -r(n-1) - l(n-1) & \end{bmatrix}$$

and has the following explicit LU factorization

$$L = \begin{bmatrix} -1 & & & \\ 1 & \ddots & & \\ & \ddots & -1 & \\ & & 1 & -1 \end{bmatrix}, \quad U = \begin{bmatrix} r(0) & -l(1) & & \\ & r(1) & \ddots & \\ & & \ddots & -l(n-1) \\ & & & r(n-1) \end{bmatrix}.$$

To find the stationary distribution we solve, in succession, the triangular systems

$$Lz_n = -e_n$$
$$Uf_n = z_n.$$

Since L is unit lower triangular we see that $z_n = e_n$ (indeed, in general one need only know U) and f_n is just the solution of $Uf_n = e_n$. Finally, backward substitution yields the known solution of equation (3).

Note that this formula does not give the values of the stationary distribution since it involves the unknown scaling factor $p(0)$. However, it does allow us to determine exactly the shape of the stationary distribution for values up to n.

4. Populations with multiple births

Now consider populations in which multiple births can occur. Although one approach to problems of this type is to re-scale the birth rate, $r(n)$, in some appropriate way and solve the single-step problem, it is not difficult to derive the solution using traditional methods. We show that, as before, the matrix approach gives the formal solution when solved analytically.

Let $r(n, k)$ be the rate at which births of k individuals occur given that the population size is n. Equation (2) becomes

$$0 = -\left(\sum_{k=1}^{\infty} r(n,k) + l(n)\right)p(n) + \sum_{k=1}^{n} r(n-k,k)p(n-k) + l(n+1)p(n+1). \quad (7)$$

The infinitesimal generator is a semi-infinite lower Hessenberg matrix whose elements are given by

$$Q^T_{i,j} = \begin{cases} -(\sum_{k=1}^{\infty} r(i,k) + l(i)) & \text{if } i = j \\ r(j, i-j) & \text{if } i > j \\ l(j) & \text{if } i+1 = j \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

The truncated system has the same form as before except that Q^T is now lower Hessenberg instead of tridiagonal. In particular, letting $p_n = l(n)p(n)f_n$ we have

$$Q^T_n f_n = -e_n. \quad (9)$$

Clearly, we need only know U from the LU factorization of Q^T to find f_n. The key to the lower Hessenberg case (which includes the tridiagonal case) is that the right-hand side in the truncated system is a rank-1 matrix (i.e. a single column).

5. Processes with multiple births and multiple deaths

In the cases considered so far, the infinitesimal generator matrix is lower Hessenberg and the solution algorithms are equivalent to the classical *solution by recursion* algorithm for finding truncated solutions. The derivations we have shown are different from the classical approach but are useful because they will allow us to look at more general birth and death models in a natural way. We now consider a process in which both multiple births and multiple deaths are allowed. We will assume that jumps as large as $\pm K$ occur in both directions (in what follows it is straightforward to extend to processes where the maximum possible number of births is not the same as the maximum possible number of deaths). The stochastic balance equation is

$$0 = \sum_{k=1}^{K} \{-(r(n,k) + l(n,k))p(n) + r(n-k,k)p(n-k) + l(n+k,k)p(n+k)\}. \tag{10}$$

The infinitesimal generator matrix is banded, with half-bandwidth K, and takes the form

$$Q^T_{i,j} = \begin{cases} -\sum_{k=1}^{K}(r(i,k) + l(i,k)) & \text{if } i = j \\ r(j, i-j) & \text{if } i > j \\ l(j, j-i) & \text{if } i < j \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The truncated system can be written

$$Q^T_n \boldsymbol{p}_n + S^{(n)} \boldsymbol{p}^{(n)} = \mathbf{0},$$

where $\boldsymbol{p}^{(n)} = [p(n)\ p(n+1)\ \ldots\ p(n+K-1)]^T$ and $S^{(n)} \in \mathbb{R}^{n \times K}$ has elements

$$S^{(n)}_{i,j} = \begin{cases} l(j, j+n-i) & \text{if } i \geq n-K+j-1 \\ 0 & \text{otherwise.} \end{cases}$$

We shall call $S^{(n)}$ the *homogeneous complement* of Q^T_n in Q^T.

Rearranging the truncated system yields

$$Q^T_n \boldsymbol{p}_n = -S^{(n)} \boldsymbol{p}^{(n)}.$$

Now, let F be the solution to

$$Q^T_n F = -S^{(n)}.$$

It follows that

$$\boldsymbol{p}_n = F \boldsymbol{p}^{(n)}, \tag{12}$$

and we see that $\boldsymbol{p}_n$ is in the range of F. If $\operatorname{rank}(F) = 1$ then we have found $\boldsymbol{p}_n$ up to an unknown scaling. This is the essence of solution by recursion since in those cases F surely has rank 1. Notice that F exists if and only if Q^T_n is non-singular; in that case $F = -Q^{-T}_n S^{(n)}$ and it is clear that $\operatorname{rank}(F) = \operatorname{rank}(S^{(n)})$.

This leads us to consider the singular value decomposition (SVD) of F; a standard approach to rank-estimation. Briefly, the SVD of a matrix $F \in \mathbb{R}^{n \times m}$ is a factorization of the form

$$F = U \Sigma V^T,$$

where both $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal, and $\Sigma \in \mathbb{R}^{n\times m}$ is diagonal. The columns $\boldsymbol{u}_j$ of U are called *left singular vectors*, the columns $\boldsymbol{v}_j$ of V are called *right singular vectors*, and the diagonal entries $\sigma_1, \sigma_2, \ldots, \sigma_m$ of Σ are called *singular values*. The singular values of a matrix are real and non-negative and are assumed to be ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$.

It is an important fact that if $\text{rank}(F) = r$ then the SVD of F has exactly r non-zero singular values and the left singular vectors, $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}$, form an orthogonal basis for the range of F.

Replacing F in equation (12) with its SVD yields

$$\boldsymbol{p}_n = U\Sigma V^T \boldsymbol{p}^{(n)};$$

which, if $\text{rank}(F) = r$, can be rewritten in the form

$$\boldsymbol{p}_n = \sum_{i=1}^{r} \boldsymbol{u}_i \sigma_i (\boldsymbol{v}_i^T \boldsymbol{p}^{(n)}). \tag{13}$$

There are several possible approaches at this point but we propose the estimate $\boldsymbol{p}_n = \alpha F\boldsymbol{x}$ where $\boldsymbol{x}$ is the solution to

$$\max_{\|\boldsymbol{x}\|_2=1} \|F\boldsymbol{x}\|_2,$$

and α is a normalization parameter. This approach has much in common with the method known as principal components analysis. It is well known, and apparent from (13) that $\boldsymbol{x} = \boldsymbol{v}_1$ is the solution to this problem and hence that our estimate of $\boldsymbol{p}_n$ is

$$\boldsymbol{p}_n = \frac{\boldsymbol{u}_1}{\boldsymbol{u}_1^T \boldsymbol{e}}.$$

It is worthwhile to note that this approach implies that $\boldsymbol{v}_1$ is a normalized estimate of $\boldsymbol{p}^{(n)}$ and hence, if $\boldsymbol{v}_1$ contains all non-negative entries this is an indication that the estimate is based on reasonable assumptions. If, however, $\boldsymbol{v}_1$ contains any negative entries then there is reason to believe that this is not a good estimate and we might consider trying a larger value of n.

Based on the preceding analysis, we propose the following algorithm to find $p(0), p(1), \ldots p(N_0)$ for a birth and death process that can have from 1 to K deaths in each time step. We assume that $N_0 > K$.

1. Let $n = N_0$.
2. Using the truncated infinitesimal generator Q_n^T and its homogeneous complement $S^{(n)}$ solve $Q_n^T F_n = -S^{(n)}$.
3. Compute the singular value decomposition of F_n, that is $U\Sigma V^T = F_n$.
4. Construct the approximation $\boldsymbol{p}_n \approx \alpha \boldsymbol{u}_1$ where α is some unknown constant and $\boldsymbol{u}_1$ is the first column of U.
5. If $\sigma_1^{(n)}$, the largest singular value of F_n, is sufficiently greater than $\sigma_2^{(n)}$ then stop and accept the approximation. Otherwise set $n := n + 1$ and return to step 2.

A key step in this algorithm is computing the singular value decomposition of F_n, where $F_n = -Q_n^{-T} S^{(n)}$. It sometimes happens that Q_n is remarkably ill-conditioned (as it approaches singularity) and hence the computation of F_n may be numerically challenging. Fortunately, it is possible to completely bypass the computation of F_n by using the generalized singular value decomposition. The form of this decomposition is outlined in the following theorem [1, 8].

Theorem. *(Generalized SVD.) If $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ and $B \in \mathbb{R}^{p\times n}$ then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$, and an invertible matrix $X \in \mathbb{R}^{n\times n}$ such that*

$$U^T AX = C,$$
$$V^T BX = S,$$

where both $C \in \mathbb{R}^{m\times n}$ and $S \in \mathbb{R}^{p\times n}$ are diagonal matrices (not necessarily square). This decomposition is known as the generalized singular value decomposition (GSVD) of (A, B).

The utility of this decomposition for our problem is as follows. If B is square and invertible then so is S and hence we can write

$$U^T AX(V^T BX)^{-1} = CS^{-1},$$

which leads to

$$U^T AB^{-1} V = CS^{-1},$$

from which we get

$$AB^{-1} = U(CS^{-1})V^T.$$

In other words, the generalized SVD of (A, B) gives us all of the components of the SVD of AB^{-1} without needing to perform the inversion (i.e. solving a linear system). For our problem, it can be shown that given the GSVD of $(-Q_n, (S^{(n)})^T)$ then

$$-Q_n^{-T} S^{(n)} = V(S^{-T} C^T)U^T,$$

so that we can use the GSVD to compute the SVD of $-Q_n^{-T} S^{(n)}$ without performing the inversion and multiplication (i.e. solving a linear system). In those cases where Q_n is ill-conditioned, this approach can help us to avoid many of the numerical difficulties.

6. A brief example

As an example we look at a simple birth-and-death process in which as many as two births or deaths can occur in the fundamental time unit. The specific model we will look at is an extension of a common insect population model characterized by the following transition rates

$$r(n, 1) = 0.237(n + 1)\,\mathrm{e}^{-0.0165n},$$
$$r(n, 2) = 0.105(n + 1)\,\mathrm{e}^{-0.0165n},$$
$$l(n, 1) = 0.088n,$$
$$l(n, 2) = 0.018n.$$

Note that the birth rate remains positive even when the population size is zero due to *migration*. The exponential factor in the birth rate function represents cannibalism of pupae by adult insects.

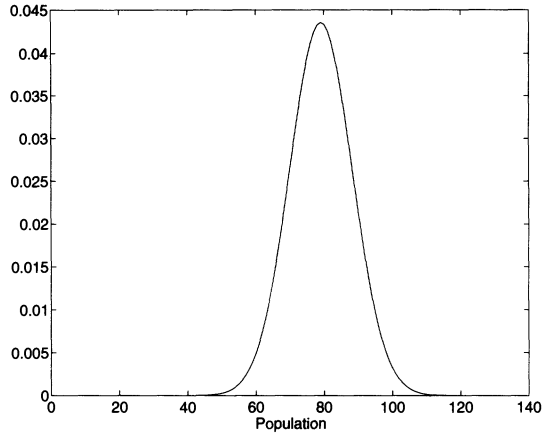

FIGURE 1: A plot of the first 140 elements of the approximate stationary distribution. The vector has been normalized so that $\boldsymbol{p}^T\boldsymbol{e} = 1$.

To compute the stationary distribution of the process analytically we would need to find the null-vector of the full infinitesimal generator Q, a semi-infinite matrix. Since this problem is generally intractable it is common to approximate the stationary distribution by looking at the eigenvector associated with the smallest-magnitude eigenvalue of Q_n, the truncated infinitesimal generator. Under proper conditions, this eigenvector will converge to the stationary distribution, as $n \to \infty$. Because of the excellent computational properties of the SVD, it is even better to use the right singular vector of Q_n associated with the smallest singular value. (All experimental calculations for this paper were made using MATLAB 4.0.) As Q_n approaches rank deficiency this singular vector gives a good approximation to the stationary distribution. Furthermore, the onset of rank deficiency can best be detected by examining the magnitudes of the small singular values which can be computed to high precision. This is not the case with small eigenvalues, which can be of low accuracy due to accumulated roundoff errors.

For this example, when $n = 250$ the smallest singular value of Q_{250} is approximately 2.486×10^{-15} and we see good convergence of the singular vector (i.e. there is very little change in the shape of the distribution as we increase n). Figure 1 contains a plot of the first 140 elements of the singular vector which we shall henceforth denote as $\boldsymbol{p}$.

We with now explore the application of the methods described in this paper. First of all, working with the matrix truncated at the first 100 equations we take the homogeneous complement, solve and return the left singular vector associated with the largest-magnitude singular value. The ratio of the largest to the second-largest singular value is roughly 2.5651×10^3. To determine how well the maximal singular vector $\boldsymbol{v}$ we have just computed approximates the shape of $\boldsymbol{p}$, it is appropriate to normalize it so as to minimize $\|\boldsymbol{p} - \alpha\boldsymbol{v}\|_2$. Solving this least-squares problem yields $\alpha = \boldsymbol{p}^T\boldsymbol{v}/\boldsymbol{v}^T\boldsymbol{v}$ and so we set $\hat{\boldsymbol{p}} = \alpha\boldsymbol{v}$. We shall call this process *shape normalization*. Figure 2 contains a plot of the logarithm of the absolute difference between the approximations after shape normalization.

Notice that the difference between the two is extremely small. It is most pronounced at the truncation point, but it is only 3.6777×10^{-5} at the worst point. Moreover, the relative error in the approximation for all n such that $p(n) > 0.000\,01$ never exceeds 1.13×10^{-2}.

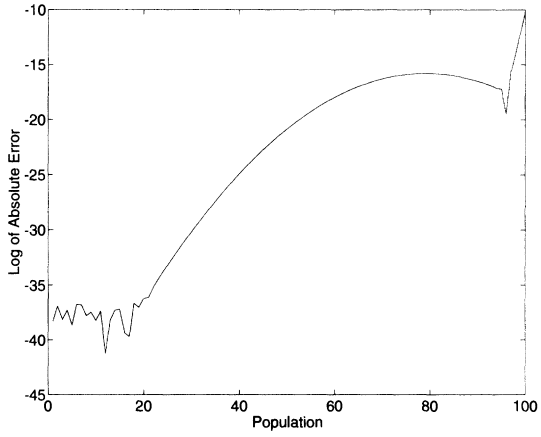

FIGURE 2: A plot of the log of the absolute error $|\boldsymbol{p} - \hat{\boldsymbol{p}}|$.

If we truncate the infinitesimal generator so that it covers population sizes ranging from 50 to 100 only, then, working with the homogeneous complement, the singular values of F are roughly $\sigma_1 = 102.2727, \sigma_2 = 5.7208, \sigma_3 = 0.0678$, and $\sigma_4 = 0.0245$. The ratio of σ_1 to σ_2 is roughly 17.8774. We can see that the method works quite well in this case, as can be seen in Figures 3 and 4.

It is interesting to try to determine how well different truncated approximations fit the stationary distribution. In order to analyse this consider the following experiment. Given $\hat{\boldsymbol{p}}$, a shape-normalized approximation to the stationary distribution for populations ranging from 0 to n, define the relative error of the approximation in the following way:

$$\frac{\sum_{i=0}^{n} | \boldsymbol{p}(i) - \hat{\boldsymbol{p}}_n |}{\sum_{i=0}^{n} \boldsymbol{p}(i)}.$$

We will consider three different truncated approximations to the stationary distribution: the minimal eigenvector of Q_n^T, the minimal right singular vector of Q_n^T, and the approximation based on the homogeneous complement method described in this paper. For each population size from $n = 1, 2, \ldots, 120$ we compute all three approximations to the stationary distribution (in the range 0 to n) and then shape normalize them. In Figure 5 we plot the relative errors of the approximations versus truncation length.

7. Generalization to quasi-birth–death processes

The method developed above can be put into a much more general framework. We begin by noting that any Markov process whose infinitesimal generator is banded is a quasi-birth–death (QBD) process. Provided we choose the block size correctly (it must be no less than the bandwidth) the infinitesimal generator can be written as a block tridiagonal matrix:

$$Q^T = \begin{bmatrix} D_0 & A_1 & & \\ B_1 & D_1 & A_2 & \\ & \ddots & \ddots & \ddots \end{bmatrix}.$$

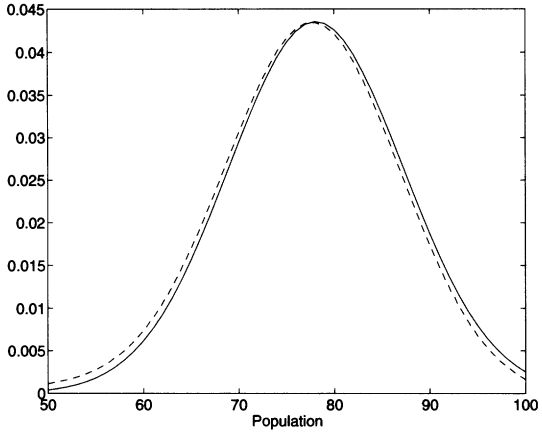

FIGURE 3: A comparison of the estimated stationary distribution in the segment 50 to 100 (dashed line) with $\boldsymbol{p}$ in the same region (solid line).

If we partition $\boldsymbol{p} = [\pi_0\ \pi_1\ \ldots]^T$ so that it is compatible for block multiplication, then the stochastic balance equations can be written in the following form:

$$D_0\pi_0 + A_1\pi_1 = \mathbf{0}$$

and

$$B_i\pi_{i-1} + D_i\pi_i + A_{i+1}\pi_{i+1} = \mathbf{0},$$

for $i = 1, 2, \ldots$.

There are two common approaches to problems of this type. First, if the Markov process represented by this matrix is *nearly completely decomposable* (NCD), that is, if the off-diagonal blocks are sufficiently small in some sense, one can disregard them and assume that the Markov chain is completely decomposable (some justification for this can be found in [5]). Then π_i are solutions for $D_i\pi_i = \mathbf{0}$ and can be solved individually. This yields approximations to the segments of $\boldsymbol{p}$, which must then be carefully assembled to get the stationary distribution (see [6] for a beautiful treatment of these methods). The problem with this approach is that it is hard to tell how good the approximations for the π_i are, since it is not clear what effect disregarding the A_i and B_i will have on the solution.

A second approach is called stochastic complementation [4]. This method also computes the segments π_i, but does so exactly using block Gaussian elimination. This method does not throw anything away so it is exact (at least on paper), but unfortunately, it is quite costly.

Our method can be extended quite naturally to this more general framework. In particular, given the block tridiagonal matrix Q^T shown above, we define the homogeneous complement of D_i in Q^T to be $S_i = [\hat{B}_{i-1}\ \hat{A}_i]$, where $\hat{A}_i$ is a matrix composed of only the non-zero columns of A_i. $\hat{B}_i$ is defined similarly, and B_0 is taken to be a zero matrix. We then solve $D_iF_i = -S_i$ and approximate π_i, with the left singular vector associated with the largest singular value of F_i. This is a simple process and allows us to estimate the quality of our approximations by examining the ratios of the largest and second-largest singular values of each F_i. This method is embarrassingly parallel.

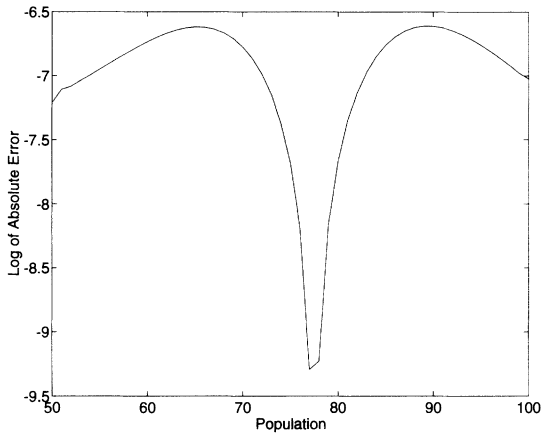

FIGURE 4: A plot of the log of the absolute error of the comparison in Figure 3.

We believe this method can also be used to generate starting guesses for those algorithms known generally as *iterative aggregation/disaggregation* (IAD). These are efficient multi-grid-like methods and include the well-known KMS [3] and Takahashi [7] algorithms. We can also develop an adaptive variation by using the algorithm described in Section 5 to solve for π_0 first. We then apply this same algorithm to solve for each successive segment using the more general definition of the homogeneous complement. This allows us to vary the block sizes adaptively so that each estimate of a segment will be a good one. This approach lacks parallelism but would give both an initial guess and a partitioning for IAD algorithms.

8. An application to NCD Markov chains

To demonstrate the general utility of the methods outlined here we will show how they can be applied to the multilevel aggregation method for nearly completely decomposable matrices. We consider an example from [6, pp. 288–294] which looks at the following NCD matrix from [2]:

$$P^T = \left[\begin{array}{ccc|ccc|ccc} 0.85 & 0.1 & 0.1 & 0 & 0.0005 & 0 & 0.00003 & 0 \\ 0 & 0.65 & 0.8 & 0.0004 & 0 & 0.00005 & 0 & 0.00005 \\ 0.149 & 0.249 & 0.0996 & 0 & 0.0004 & 0 & 0.00003 & 0 \\ \hline 0.0009 & 0 & 0.0003 & 0.7 & 0.399 & 0 & 0.00004 & 0 \\ 0 & 0.0009 & 0 & 0.2995 & 0.6 & 0.00005 & 0 & 0.00005 \\ \hline 0.00005 & 0.00005 & 0 & 0 & 0.0001 & 0.6 & 0.1 & 0.1999 \\ 0 & 0 & 0.0001 & 0.0001 & 0 & 0.2499 & 0.8 & 0.25 \\ 0.00005 & 0.00005 & 0 & 0 & 0 & 0.15 & 0.0999 & 0.55 \end{array}\right].$$

Following the indicated partitioning we consider the block matrix

$$P^T = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}.$$

The fundamental steps in the algorithm are:

1. Approximate the probability distribution of the ith block with $\boldsymbol{u}_i$, where

$$P_{ii}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \quad \text{and} \quad \boldsymbol{u}_i^T \boldsymbol{e} = 1,$$

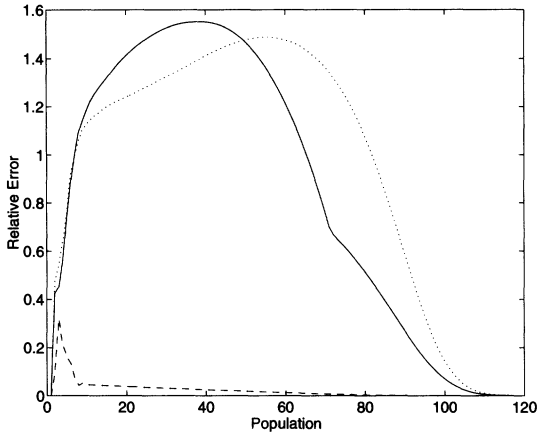

FIGURE 5: The relative error of three truncated approximations: the minimal eigenvector (solid line), the minimal right singular vector (dotted line), and the homogeneous complement method from this paper (dashed line).

where λ_i is the Perron root of P_{ii}.

2. Approximate the block aggregation matrix A by

$$\tilde{A} = \begin{bmatrix} \boldsymbol{e}^T & 0 & 0 \\ 0 & \boldsymbol{e}^T & 0 \\ 0 & 0 & \boldsymbol{e}^T \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & 0 & 0 \\ 0 & \boldsymbol{u}_2 & 0 \\ 0 & 0 & \boldsymbol{u}_3 \end{bmatrix}.$$

3. Compute $\boldsymbol{p}$, the stationary distribution of $\tilde{A}$,

$$\tilde{A}\boldsymbol{p} = \lambda \boldsymbol{p} \quad \text{and} \quad \boldsymbol{p}^T\boldsymbol{e} = 1,$$

where λ is the Perron root of $\tilde{A}$.

4. Approximate the stationary distribution of P with

$$\begin{bmatrix} p_1\boldsymbol{u}_1 \\ p_2\boldsymbol{u}_2 \\ p_3\boldsymbol{u}_3 \end{bmatrix}.$$

To integrate the techniques we have outlined into this algorithm we modify step 1 to use homogeneous complementation. In particular, we let $\boldsymbol{u}_1$ be the left singular vector associated with the largest singular value of

$$(P_{11} - I)^{-1}[P_{12}\ P_{13}].$$

Similarly, we let $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ be the left singular vectors associated with the largest singular values of

$$(P_{22} - I)^{-1}[P_{21}\ P_{23}] \quad \text{and} \quad (P_{33} - I)^{-1}[P_{31}\ P_{32}],$$

respectively.

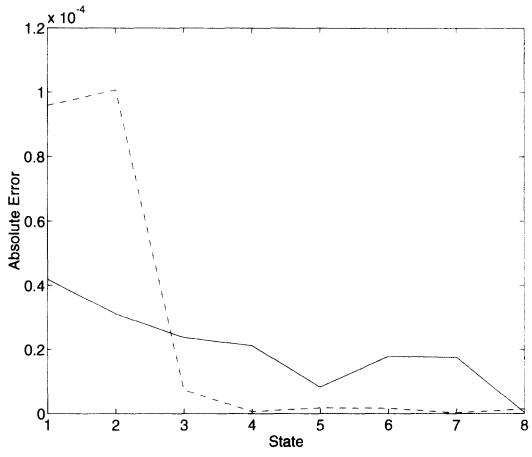

FIGURE 6: The solid line shows the absolute error in approximating the stationary distribution of the Courtois matrix using the method described in [6], the dashed line is the absolute error of our modification of this method.

If we compute the stationary distribution of P we find that

$$\pi = [\,0.08928 \quad 0.09276 \quad 0.04049 \quad 0.15853 \quad 0.11894 \quad 0.12039 \quad 0.27780 \quad 0.10182\,].$$

The approximate stationary distribution using the standard method as described in [6] is

$$\tilde{\pi} = [\,0.08932 \quad 0.09273 \quad 0.04046 \quad 0.15855 \quad 0.11893 \quad 0.12037 \quad 0.27781 \quad 0.10182\,].$$

Finally, the approximate stationary distribution using the method just described is

$$\hat{\pi} = [\,0.08938 \quad 0.09266 \quad 0.04050 \quad 0.15853 \quad 0.11894 \quad 0.12038 \quad 0.27780 \quad 0.10182\,].$$

In Figure 6 we plot the absolute errors of the two approximations. Note that both methods have errors of comparable magnitude (10^{-4}).

Now consider what happens if we modify the Courtois matrix with a parameterized perturbation as shown below

$$P^T(\epsilon) = \left[\begin{array}{ccc|ccc|ccc} 0.85 & 0.1 & 0.1 & 0 & 0.0005 & 0 & 0.00003 & 0 \\ 0 & 0.65-\epsilon & 0.8-\epsilon & 0.0004 & \epsilon & 0.00005 & \epsilon & 0.00005 \\ 0.149 & 0.249 & 0.0996 & 0 & 0.0004 & 0 & 0.00003 & 0 \\ \hline 0.0009 & \epsilon & 0.0003 & 0.7-\epsilon & 0.399-\epsilon & \epsilon & 0.00004 & 0 \\ 0 & 0.0009 & 0 & 0.2995 & 0.6 & 0.00005 & 0 & 0.00005 \\ \hline 0.00005 & 0.00005 & \epsilon & \epsilon & 0.0001 & 0.6-\epsilon & 0.1 & 0.1999 \\ 0 & 0 & 0.0001 & 0.0001 & 0 & 0.2499 & 0.8-\epsilon & 0.25 \\ 0.00005 & 0.00005 & 0 & 0 & 0 & 0.15 & 0.0999 & 0.55 \end{array}\right].$$

If we allow ϵ to vary over the interval $[0, 0.399]$ (we can do this without violating the stochasticity of the matrix) we can compare how well the two methods approximate the stationary distribution of $P(\epsilon)$ over the range of ϵ. Figure 7 shows a log plot of the errors where

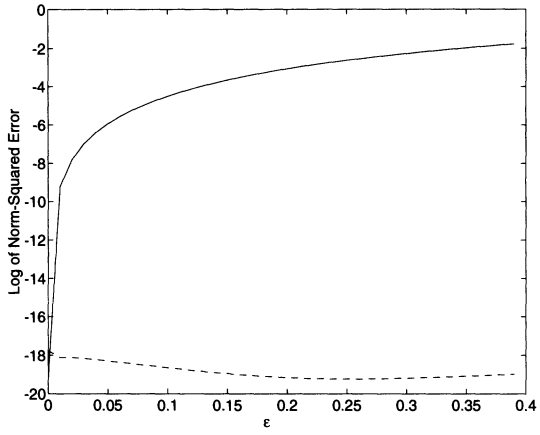

FIGURE 7: Log plot of the error of approximation for the parameterized Courtois matrix. The solid line shows $\log \|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|^2$ and the dashed line shows $\log \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|^2$.

one can see that the second approximation is quite good over the entire range of ϵ while the standard one degrades as the magnitude of the perturbation grows.

Of course, we expect growth in the error of the first approximation because as ϵ grows, the matrix begins to violate the NCD assumption. However, the second approximation continues to work well because the homogeneous complement approach takes into account the structure of the off-diagonal blocks as well as their magnitude. Note that the homogeneous complement of the P_{11} block in this matrix is

$$[P_{12}P_{13}] = \left[\begin{array}{cc|ccc} 0 & 0.0005 & 0 & 0.00003 & 0 \\ 0.0004 & \epsilon & 0.000\,05 & \epsilon & 0.000\,05 \\ 0 & 0.0004 & 0 & 0.00003 & 0 \end{array}\right].$$

As ϵ grows this matrix approaches a rank-1 matrix. In fact, when $\epsilon = 0.399$ the singular values of this matrix are $\sigma_1 = 0.5643$, $\sigma_2 = 0.4230 \times 10^{-3}$, and $\sigma_3 = 0.5106 \times 10^{-8}$, so that although it is full-rank, it is quite close to a rank-1 matrix. Since we can think of the homogeneous complement as a *coupling* to the other states in the chain, we call a situation like this a *weak geometric coupling*. We might also call it a *low-rank coupling* except that in this, and most such cases, the homogeneous complement does have full rank. The term *weak geometric coupling* expresses the fact that the real problem here is that the columns of the homogeneous complement are geometrically quite nearby. Note that for this example the homogeneous complements of all blocks exhibit weak geometric coupling as ϵ gets large.

Of course, when $\epsilon = 0$ our method still works well even though none of the homogeneous complements exhibit weak geometric coupling in that case. Indeed, it is possible to gain some insight into why this should generally be the case for an NCD matrix. We begin by noting that if the matrix is NCD then each diagonal block, P_{ii}, will be very nearly stochastic and so will have a Perron root λ_1 such that $0 < 1 - \lambda_1 \ll 1$. It follows that $-(P_{ii} - I)^{-1}$ exists and has a maximal eigenvalue $\mu_1 = (1 - \lambda_1)^{-1} \gg 1$. Moreover, if λ_2, the second-largest eigenvalue of P_{ii}, is sufficiently well separated from λ_1 then μ_1 will be much larger than any of the remaining eigenvalues of $-(P_{ii} - I)^{-1}$.

Now let S_{ii} be the homogeneous complement of P_{ii} and denote its jth column by $\boldsymbol{s}_j$. For each column of S_{ii} we can write

$$\boldsymbol{s}_j = \alpha_j \boldsymbol{v}_1 + \boldsymbol{c}_j,$$

where $\boldsymbol{v}_1$ is the Perron vector of $-(P_{ii} - I)^{-1}$ (and hence also of P_{ii}) and $\boldsymbol{c}_j$ is the projection of $\boldsymbol{s}_j$ onto the orthogonal complement of $\boldsymbol{v}_1$. Now we can write

$$S_{ii} = \boldsymbol{v}_1[\alpha_1 \ \alpha_2 \ \ldots] + C,$$

where $C = [\boldsymbol{c}_1 \ \boldsymbol{c}_2 \ \ldots]$. From which it now follows that

$$-(P_{ii} - I)^{-1} S_{ii} = \mu_1 \boldsymbol{v}_1[\alpha_1 \ \alpha_2 \ \ldots] - (P_{ii} - I)^{-1} C.$$

Now, since the columns of C are in the orthogonal complement of $\boldsymbol{v}_1$ it follows that the elements of $(P_{ii} - I)^{-1} C$ can have magnitude no greater than $\mu_2 \max |C_{ij}|$. So, if μ_1 is sufficiently larger than μ_2 then we would expect that the most dominant component of the columns of $-(P_{ii} - I)^{-1} S_{ii}$ will be close to $\boldsymbol{v}_1$. If this is true then the left singular vector of $-(P_{ii} - I)^{-1} S_{ii}$ associated with the largest singular value will also be close to $\boldsymbol{v}_1$. And we see that we can reasonably expect that our method will return a vector that is quite close to that returned by the standard algorithm in this case.

In fact, if μ_1 is sufficiently dominant then $-(P_{ii} - I)^{-1} S_{ii}$ will tend to approach the rank-1 matrix $\mu_1 \boldsymbol{v}_1[\alpha_1 \ \alpha_2 \ \ldots]$, which implies that the approximation will become exact. This is a simple way of seeing why the standard method works.

Acknowledgements

The authors would like to thank the reviewers for a number of insightful suggestions that helped us to refine our understanding of the problem and have significantly improved the presentation. The first author gratefully acknowledges support by direct grant from the Naval Postgraduate School.

References

[1] BAI, Z. AND DEMMEL, J. W. (1993). Computing the generalized singular value decomposition. *SIAM J. Scientific Computing* **14,** 1464–1486.

[2] COURTOIS, P. J. (1977). *Decomposability: Queueing and Computer System Applications*. Academic Press, New York.

[3] KOURY, R., MCALLISTER, D. F. AND STEWART, W. J. (1984). Methods for computing the stationary distributions of nearly-completely-decomposable Markov chains. *SIAM J. Algebraic Discrete Mathematics* **5,** 164–186.

[4] MEYER, C. D. (1989). Stochastic complementation, uncoupling Markov chains and the theory of nearly reducible systems. *SIAM Review* **31,** 240–272.

[5] SIMON, H. A. AND ANDO, A. (1961). Aggregation of variables in dynamic systems. *Econometrica* **29,** 111–138.

[6] STEWART, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.

[7] TAKAHASHI, Y. (1975). A Lumping Method for Numerical Calculation of Stationary Distributions of Markov Chains. Technical Report B-18, Department of Information Sciences, Tokyo Institute of Technology, Tokyo, Japan.

VAN LOAN, C. F. (1976). Generalizing the singular value decomposition. *SIAM J. Numerical Analysis* **13,** 76–83.