Res. Lett. Inf. Math. Sci., 2005, Vol.**8**, pp155-163
Available online at http://iims.massey.ac.nz/research/letters/

On a matrix with integer eigenvalues and its relation to conditional Poisson sampling

L. Bondesson[1] & I. Traat[2]

[1]*Department of Mathematics and Mathematical Statistics*
University of Umeå, Sweden

[2]*Institute of Mathematical Statistics*
University of Tartu, Estonia

A special non-symmetric $N \times N$ matrix with eigenvalues $0, 1, 2, \ldots, N-1$ is presented. The matrix appears in sampling theory. Its right eigenvectors, if properly normalized, give the inclusion probabilities of the Conditional Poisson design (for all different fixed sample sizes). The explicit expressions for the right eigenvectors become complicated for N large. Nevertheless, the left eigenvectors have a simple analytic form. An inversion of the left eigenvector matrix produces the right eigenvectors − the inclusion probabilities. Finally, a more general matrix with similar properties is defined and expressions for its left and right eigenvectors are derived.

1 Introduction

There are many real $N \times N$ matrices $\mathbf{A}$ that have eigenvalues $0, 1, 2, \ldots, N-1$. The most general one is given by $\mathbf{A} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{-1}$, where $\mathbf{P}$ is a matrix with linearly independent column vectors and $\boldsymbol{\Delta}$ is a diagonal matrix with the given integers on the diagonal; cf., e.g., Eves (1980, pp. 223-225). However, in this paper we present a special non-symmetric matrix $\mathbf{A}$ with these eigenvalues. The matrix $\mathbf{A}$ is simple but not the matrix $\mathbf{P}$ of its right eigenvectors. The right eigenvectors of $\mathbf{A}$ are of special interest in sampling theory. They give inclusion probabilities of an important sampling design − the Conditional Poisson (CP) design. They are difficult to calculate in general and simple explicit analytic formulae are not available. In the literature (Chen, 1997, Aires, 1999, Traat *et al.*, 2004, Bondesson *et al.*, 2004), some recursive formulae and algorithms for their calculation are presented. In this

paper we develop an eigenvector approach. The eigenvector, properly normalized, corresponding to the eigenvalue $n-1$ ($n = 1, 2, \ldots, N$) represents the inclusion probabilities of the CP design with fixed sample size n. Consequently, the matrix $\mathbf{P}$ yields the inclusion probabilities for all possible sample sizes, simultaneously.

Surprisingly, in spite of the complexity of $\mathbf{P}$, the matrix $\mathbf{Q} = \mathbf{P}^{-1}$ of left eigenvectors has a simple form. The matrix $\mathbf{Q}$ is presented in this paper. The inclusion probabilities of the CP design can now be obtained by just inverting $\mathbf{Q}$.

It appears that the matrix $\mathbf{A}$ is a special case of a more general matrix that might have interest in pure linear algebra as well. Expressions for the left and right eigenvectors of this matrix are derived.

The paper is organized as follows. First, the sampling origin of the matrix $\mathbf{A}$ is given. The matrix appears as the coefficients in a linear equation system with the inclusion probabilities of the CP design (of size n) as unknowns. The solution of the system is an eigenvector of $\mathbf{A}$ that corresponds to the eigenvalue $n-1$. Expressions for the right and left eigenvectors are presented.

In the second part of the paper, which also contains proofs, generalizations are made to a wider class of matrices, still having consecutive integer eigenvalues and the same simple expressions for the left eigenvectors. Explicit expressions for the right eigenvectors are given as well. The case of equal elements is treated as a limiting case. Left eigenvectors are not much treated in the matrix literature. This paper demonstrates their usefulness in practical problems involving non-symmetric matrices.

2 The sampling origin

Here we describe how the matrix $\mathbf{A}$ (not yet defined) appears in sampling theory. Consider a finite population with N units, $U = \{1, 2, \ldots, N\}$. Let $\mathbf{I}$ be a multivariate random Bernoulli vector describing random selection of units from U:

$$\mathbf{I} = (I_1, I_2, \ldots, I_N), \quad \Pr(I_i = 1) = p_i, \quad I_i \text{ independent.}$$

An outcome of $\mathbf{I}$ is a Poisson sample; $I_i = 1$ means that unit i is sampled and $I_i = 0$ means that it is not. The number p_i is the inclusion probability of unit i under the Poisson design. The sample has random size $\sum I_i$. Here and below all sums and products without range specification mean that the index runs from 1 to N.

Now, let us consider a new Bernoulli vector conditional on a sample size:

$$\mathbf{I}^{CP} = \mathbf{I} \mid \sum I_i = n),$$

where n is any fixed number between 1 and N. The outcome of $\mathbf{I}^{CP}$ is a Conditional Poisson sample. It has fixed sample size. Denote the CP inclusion probabilities by

$$\pi_i = \Pr(I_i^{CP} = 1), \quad \pi_{ij} = \Pr(I_i^{CP} = 1, I_j^{CP} = 1).$$

There exists a simple formula connecting these inclusion probabilities with the original p_is (e.g. Aires, 1999):

$$\pi_{ij} = \frac{p_j q_i}{p_j - p_i}\pi_i + \frac{p_i q_j}{p_i - p_j}\pi_j,\ i \neq j,\ p_i \neq p_j, \tag{1}$$

where $q_i = 1 - p_i$. This formula can also be proved by a Gibbs sampling reasoning as is now indicated.

Assume that a CP-sample of size n has been generated. Two units i and j are then picked out such that $I_i = 1,\ I_j = 0$ or $I_i = 0,\ I_j = 1$. Without changing the distribution, we can simulate new values of I_i and I_j by using the probabilities

$$\Pr(I_i = 1, I_j = 0) = \frac{p_i q_j}{p_i q_j + p_j q_i} \text{ and } \Pr(I_i = 0, I_j = 1) = \frac{p_j q_i}{p_i q_j + p_j q_i}. \tag{2}$$

Note that the probabilities in (2) can be expressed through inclusion probabilities, e.g. $\Pr(I_i = 1, I_j = 0) = \Pr(I_i = 1) - \Pr(I_i = 1, I_j = 1)$. Now we get under stability that the probabilities in (2) must equal $\pi_i - \pi_{ij}$ and $\pi_j - \pi_{ij}$, respectively. Dividing the two obtained equations by each other and solving for π_{ij} we get (1). Cf. Bondesson *et al.* (2004)

The following relation holds for any fixed size sampling design:

$$\sum_{j;\, j \neq i} \pi_{ij} = (n-1)\pi_i, \tag{3}$$

Using (1) and (3), we get the equation system

$$(\sum_{j;\, j \neq i} \frac{p_j q_i}{p_j - p_i})\pi_i + \sum_{j;\, j \neq i} \frac{p_i q_j}{p_i - p_j}\pi_j = (n-1)\pi_i,\ i = 1, 2, \ldots, N.$$

We want to solve it with respect to π_i. Denoting the vector of these probabilities by $\boldsymbol{\pi}$: $N \times 1$, we get the system in matrix form:

$$\mathbf{A}\boldsymbol{\pi} = (n-1)\boldsymbol{\pi},$$

where

$$\mathbf{A} = \text{diag}(\mathbf{1}^T\mathbf{C}) + \mathbf{C}:\ N \times N, \tag{4}$$

$$\text{with } c_{ij} = \frac{p_i q_j}{p_i - p_j},\quad c_{ii} = 0,\quad p_i \neq p_j, \tag{5}$$

and $\mathbf{1}$ is a column vector of ones. We see that the unknown $\boldsymbol{\pi}$ is the right eigenvector of $\mathbf{A}$ corresponding to the eigenvalue $n-1$. Since n is any integer $1, 2, \ldots, N$, the eigenvalues of $\mathbf{A}$ are consecutive integers $\{0, 1, \ldots, N-1\}$.

3 The matrix A

The matrix $\mathbf{A}$ in (4)-(5) has a simple form. It is non-symmetric but with a special structure,

$$a_{ij} = c_{ij} = 1 - c_{ji}.$$

Its elements have the alternate form,

$$c_{ij} = \frac{p_i q_j}{p_i q_j - p_j q_i}.$$

Its diagonal elements are the column sums of $\mathbf{C}$, $a_{ii} = \mathbf{1}^T\mathbf{C}$. Its rows sum up to a constant,

$$\mathbf{A1} = (N-1)\mathbf{1},$$

showing that the eigenvector corresponding to the largest eigenvalue has all its coordinates equal. In the sampling context, it means that for a sample of size $n = N$ (full population), the inclusion probabilities are equal to 1. The matrix $\mathbf{A}$ is singular, it has an eigenvalue 0.

As an example, let the p_is be given by the vector $\mathbf{p} = (6/10, 7/10, 8/10, 9/10)$. The p_is sum to 3. Then the matrix $\mathbf{A}$ is:

$$\mathbf{A} = \begin{bmatrix} \frac{28}{5} & -9/5 & -3/5 & -1/5 \\ \frac{14}{5} & \frac{39}{20} & -7/5 & -\frac{7}{20} \\ 8/5 & \frac{12}{5} & -1/5 & -4/5 \\ 6/5 & \frac{27}{20} & 9/5 & -\frac{27}{20} \end{bmatrix}.$$

Its eigenvalues are $3, 2, 1, 0$. Right and left eigenvectors are provided by the following matrices:

$$\mathbf{P} = \frac{1}{2500}\begin{bmatrix} -756 & 597 & -138 & 9 \\ -756 & 777 & -203 & 14 \\ -756 & 912 & -308 & 24 \\ -756 & 1017 & -423 & 54 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} -\frac{160}{9} & \frac{135}{7} & -5 & \frac{5}{27} \\ -\frac{80}{3} & 45 & -20 & 5/3 \\ -40 & 105 & -80 & 15 \\ -60 & 245 & -320 & 135 \end{bmatrix}.$$

The columns in $\mathbf{P}$ are right eigenvectors and the rows in $\mathbf{Q}$ are left eigenvectors. Here $\mathbf{Q} = \mathbf{P}^{-1}$. After normalizing each column of $\mathbf{P}$ to sum up to a sample size: first column to N, second to $N-1$, etc., we get the eigenvector-matrix $\mathbf{P}^*$ with inclusion probabilities for all possible sample sizes in columns:

$$\mathbf{P}^* = \begin{bmatrix} 1 & 0.542 & 0.257 & 0.089 \\ 1 & 0.706 & 0.379 & 0.139 \\ 1 & 0.828 & 0.575 & 0.238 \\ 1 & 0.924 & 0.789 & 0.534 \end{bmatrix}.$$

For comparison note that the Poisson design with expected sample size 3 has inclusion probabilities given by the vector $\mathbf{p}$ but the CP design with fixed sample size 3 has inclusion probabilities given in the second column of $\mathbf{P}^*$.

4 Eigenvectors of A

In sampling context, we may use directly the definition of inclusion probabilities to construct the eigenvectors. Unit i is included if any of the samples $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ in which $y_i = 1$ occurs. Here $y_j \in \{0, 1\}$ and $|\mathbf{y}| = \sum y_j = n$, the sample size. Consequently, for the CP design we have:

$$\pi_i = \frac{1}{c} \sum_{\mathbf{y};\ y_i=1,\, |\mathbf{y}|=n} \prod_{j=1}^{N} p_j^{y_j} q_j^{1-y_j},$$

where the normalizing constant c can be found from the condition $\sum \pi_j = n$. For example, for $N = 3$ we have the non-normalized eigenvectors for sample sizes $n = 3, 2, 1$:

$$\mathbf{P} = \begin{bmatrix} p_1p_2p_3 & p_1(p_2q_3 + q_2p_3) & p_1q_2q_3 \\ p_2p_1p_3 & p_2(p_1q_3 + q_1p_3) & p_2q_1q_3 \\ p_3p_1p_2 & p_3(p_1q_2 + q_1p_2) & p_3q_1q_2 \end{bmatrix}.$$

Note that here we solved the algebraic problem of finding eigenvalues-eigenvectors with the help of their statistical meaning: eigenvalues are sample sizes minus 1, and eigenvectors are the inclusion probabilities of the CP design. This adds one more example to the statistical proofs of matrix results by Rao (2000). Though the expressions in $\mathbf{P}$ are simple functions of p_j and q_j, finding all the necessary combinations may become time consuming for N large. However, it appears that the left eigenvectors have a simple formula for any N.

The matrix $\mathbf{Q}$ of left eigenvectors has elements in the form

$$Q_{ij} = p_j^{i-2} q_j^{N-i} / u_j, \tag{6}$$

where

$$u_j = \prod_{k=1,\ k\neq j}^{N} (p_j q_k - p_k q_j). \tag{7}$$

One can easily check by a computer algebra package that $\mathbf{Q}$ satisfies the equation of left eigenvectors, $\mathbf{QA} = \mathbf{\Delta Q}$. A proof (in a more general setting) is given in section 5. Right eigenvectors are obtained by inverting $\mathbf{Q}$. The matrices $\mathbf{P}$ and $\mathbf{Q}$ in section 3 were calculated in this way.

5 Generalizations

5.1 The matrix A in a more general form

Let $\mathbf{A} = (a_{ij}) :\ N \times N$ be defined as

$$a_{ij} = \frac{r_i}{r_i - r_j},\ i \neq j,\ \text{and}\ a_{ii} = \sum_{i;\ i\neq j} a_{ij}, \tag{8}$$

where the r_is are distinct *real* or *complex* numbers. This matrix reduces to the special case (4)-(5) if we set $r_i = p_i/(1-p_i)$, $0 < p_i < 1$.

Let $\mathbf{Q} = (Q_{ij}) :\ N \times N$ be defined as

$$Q_{ij} = r_j^{i-2}/u_j, \text{ where } u_j = \prod_{k;\ k\neq j} (r_j - r_k). \tag{9}$$

In the following we show that the matrix $\mathbf{A}$ has integer eigenvalues, and its left eigenvectors are given by the rows of $\mathbf{Q}$.

Theorem 1. *The eigenvalues of* $\mathbf{A}$ *are the integers* $0, 1, 2, \ldots, N-1$. *Moreover,* $\mathbf{QA} = \mathbf{\Delta Q}$, *where* $\mathbf{\Delta} = \mathrm{diag}(N-1, N-2, \ldots, 2, 1)$.

Proof: We must prove, for $i = 1, 2, \ldots, N$, $k = 1, 2, \ldots, N$, that $\sum_{j=1}^{N} Q_{ij} a_{jk} = (N-i)Q_{ik}$, i.e.

$$\sum_{j;\ j\neq k} r_j^{i-2} \Big(\prod_{\nu;\ \nu\neq j} \frac{1}{r_j - r_\nu} \Big) \frac{r_j}{r_j - r_k} + r_k^{i-2} \Big(\prod_{\nu;\ \nu\neq k} \frac{1}{r_k - r_\nu} \Big) \sum_{\mu;\ \mu\neq k} \frac{r_\mu}{r_\mu - r_k}$$

$$= (N-i) r_k^{i-2} \prod_{\nu;\ \nu\neq k} \frac{1}{r_k - r_\nu}.$$

For symmetry reasons, it suffices to consider the case where $k = N$. We put $r_N = x$ and set $N = M + 1$. After some reordering of the terms and factors and since $r/(r-x) = 1 + x/(r-x)$, the formula above transforms into

$$\sum_{j=1}^{M} r_j^{i-2} \Big(\prod_{\nu;\ \nu\neq j}^{M} \frac{1}{r_j - r_\nu} \Big) \frac{r_j}{(r_j - x)^2} = (-1)^{M-1} x^{i-2} \Big(\prod_{\nu=1}^{M} \frac{1}{r_\nu - x} \Big) \Big(i - 1 + \sum_{\mu=1}^{M} \frac{x}{r_\mu - x} \Big).$$

The left-hand side is called $L(x)$ and the right-hand side $R(x)$. Apparently $L(x)$ is of the form $\sum_{j=1}^{M} b_j (r_j - x)^{-2}$, where each b_j is a constant. As to the function $R(x)$, it is easily seen that it equals the derivative of

$$g(x) = (-1)^{M-1} x^{i-1} \prod_{\nu=1}^{M} \frac{1}{r_\nu - x},$$

which by a partial fraction decomposition can be rewritten into the form $g(x) = \sum_{j=1}^{M} c_j (r_j - x)^{-1}$. Hence $R(x) = \sum_{j=1}^{M} c_j (r_j - x)^{-2}$. Multiplying $L(x)$ and $R(x)$ by $(r_\nu - x)^2$ and then letting $x \to r_\nu$, we see that $b_\nu = c_\nu$ for each ν. Hence $L(x) = R(x)$ as desired, and the proof is completed.

The inverse of $\mathbf{Q}$ gives the right eigenvectors of $\mathbf{A}$. The explicit expression for $\mathbf{Q}^{-1}$ is given in the next theorem.

Theorem 2. *The inverse matrix* $\mathbf{Q}^{-1}$ *is given by* $\mathbf{P} = (P_{ij})$, *where*

$$P_{ij} = (-1)^{N+j} r_i \rho_{N-j}^{(-i)}, \tag{10}$$

and $\rho_{N-j}^{(-i)}$ *denotes the sum of all possible products of* $N-j$ *distinct factors from the set* $\{r_1, r_2, \ldots, r_{i-1}, r_{i+1}, \ldots, r_N\}$.

Proof: Let $\mathbf{P}$ be as stated. We can either prove that $\mathbf{QP} = \mathbf{I}$ or that $\mathbf{PQ} = \mathbf{I}$. We choose to verify the last relation, i.e.

$$\sum_{j=1}^{N} P_{ij} Q_{jk} = \delta_{ik}, \tag{11}$$

where $\delta_{ik} = 0$ if $i \neq k$ and otherwise 1. Apparently

$$\sum_{j=1}^{N} P_{ij} Q_{jk} = r_i \Big(\prod_{\nu;\, \nu \neq k}^{N} \frac{1}{r_k - r_\nu} \Big) \sum_{j=1}^{N} r_k^{j-2} (-1)^{N+j} \rho_{N-j}^{(-i)}.$$

We set $r_k = x$ and then (11) will be equivalent to

$$\sum_{j=1}^{N} r_i x^{j-2} (-1)^{N+j} \rho_{N-j}^{(-i)} = \delta_{ik} \prod_{\nu;\ \nu \neq k}^{N} (x - r_\nu).$$

For $i = k$, we have $r_i = x$ and hence $r_i x^{j-2} = x^{j-1}$. Obviously the relation holds since the left-hand side is the Taylor expansion of the product on the right-hand side. If $i \neq k$, we must show that $\sum_{j=1}^{N} x^{j-2} (-1)^{N+j} \rho_{N-j}^{(-i)} = 0$ for $x = r_k$ or equivalently that

$$\sum_{j=1}^{N} x^{j-1} (-1)^{N+j} \rho_{N-j}^{(-i)} = 0$$

for $x = r_k$. However, the left-hand side is just the Taylor expansion of $\prod_{\nu;\, \nu \neq i} (x - r_\nu)$ and the product is 0 at $x = r_k$ because the factor $x - r_k$ is not excluded. The proof is completed.

As an example we present the related matrices in the 4×4 case. The matrix itself is:

$$\mathbf{A} = \begin{bmatrix} v_1 & \frac{r_1}{r_1 - r_2} & \frac{r_1}{r_1 - r_3} & \frac{r_1}{r_1 - r_4} \\ \frac{r_2}{r_2 - r_1} & v_2 & \frac{r_2}{r_2 - r_3} & \frac{r_2}{r_2 - r_4} \\ \frac{r_3}{r_3 - r_1} & \frac{r_3}{r_3 - r_2} & v_3 & \frac{r_3}{r_3 - r_4} \\ \frac{r_4}{r_4 - r_1} & \frac{r_4}{r_4 - r_2} & \frac{r_4}{r_4 - r_3} & v_4 \end{bmatrix},$$

where $v_j = \sum_{k;\ k \neq j} \frac{r_k}{r_k - r_j}$, $j = 1, 2, 3, 4$, are the column sums. The left eigenvectors in the rows of $\mathbf{Q}$ are:

$$\mathbf{Q} = \begin{bmatrix} r_1^{-1}/u_1 & r_2^{-1}/u_2 & r_3^{-1}/u_3 & r_4^{-1}/u_4 \\ 1/u_1 & 1/u_2 & 1/u_3 & 1/u_4 \\ r_1/u_1 & r_2/u_2 & r_3/u_3 & r_4/u_4 \\ r_1^2/u_1 & r_2^2/u_2 & r_3^2/u_3 & r_4^2/u_4 \end{bmatrix},$$

where $u_j = \prod_{k;\ k\neq j}(r_j - r_k)$, $j = 1, 2, 3, 4$. The right eigenvectors in the columns of $\mathbf{P}$ are:

$$\mathbf{P} = \begin{bmatrix} -r_1r_2r_3r_4 & r_1(r_2r_3 + r_2r_4 + r_3r_4) & -r_1(r_2 + r_3 + r_4) & r_1 \\ -r_1r_2r_3r_4 & r_2(r_1r_3 + r_1r_4 + r_3r_4) & -r_2(r_1 + r_3 + r_4) & r_2 \\ -r_1r_2r_3r_4 & r_3(r_1r_2 + r_1r_4 + r_2r_4) & -r_3(r_1 + r_2 + r_4) & r_3 \\ -r_1r_2r_3r_4 & r_4(r_1r_2 + r_1r_3 + r_2r_3) & -r_4(r_1 + r_2 + r_3) & r_4 \end{bmatrix}.$$

In row i and column j of $\mathbf{P}$, one finds the signed sum of all possible products containing r_i and $N - j$ other distinct factors. The sign is $(-1)^{N+j}$.

5.2 Equal elements $\mathbf{r}_i$

The results in section 5.1 provide a method to calculate simultaneously and rapidly all the product sums displayed in $\mathbf{P}$. The simple matrix $\mathbf{Q}$ is just inverted exactly or numerically in high precision using a computer. However, it is required that all the r_is are distinct. If some r_i are equal, the matrix $\mathbf{P}$ is still well defined, though singular, but $\mathbf{A}$ and $\mathbf{Q}$ are not. Such a $\mathbf{P}$ is needed in sampling theory, where it yields inclusion probabilities for the CP design.

In the case of equal r_is, the method can be modified so that it still works. We first make all the r_is distinct by just adding a multiple of a small number x to equal elements. For example, if $r_i = r_{i+1} = \ldots = r_{i+\nu}$, then the modified elements are $r'_i = r_i,\ r'_{i+1} = r_i + x,\ \ldots\ ,\ r'_{i+\nu} = r_i + \nu x$. For the modified elements the matrix $\mathbf{Q}$ exists and can be calculated by (9). Next we invert $\mathbf{Q}$ and get $\mathbf{P}$ for the r'_is. As x tends to 0, $\mathbf{P}$ tends to the desired matrix with the original r_is.

Of course, there are also obvious recursive methods to obtain the product sums in $\mathbf{P}$, both for distinct and non distinct r_is. For example, consider the function $g_i(x) = \prod_{\nu;\,\nu\neq i}(x + r_\nu)$. By a Taylor expansion, we have

$$g_i(x) = x^{N-1} + \rho_1^{(-i)}x^{N-2} + \rho_2^{(-i)}x^{N-3} + \ldots.$$

The coefficient $\rho_{j-1}^{(-i)}$ in front of x^{N-j} gives the sum of all products of $j - 1$ distinct factors, all different from r_i. Computer algebra packages are able to calculate high order Taylor expansions. Thus, except for the sign, row i in $\mathbf{P}$ is obtained by just reading off all the Taylor coefficients in reversed order and then multiplying by r_i. Certainly this method is more stable than the inversion method for large values of N.

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodol. Comput. Appl. Probab.* **4**, 457-469.

Bondesson, L., Traat, I., Lundqvist, A. (2004). Pareto sampling versus Sampford and Conditional Poisson sampling. *Research Report No. 6*, Department of Mathematical Statistics, Umeå University.

Chen, S.X., Liu, J.S. (1997). Statistical applications to the Poisson-binomial and conditional Bernoulli distributions. *Statist. Sinica* **7**, 875-892.

Eves, H. (1980). *Elementary Matrix Theory.* Dover, New York.

Rao, C.R. (2000). Statistical proofs of some matrix inequalities. *Linear Algebra Appl.* **321**, 307-320.

Traat, I., Bondesson, L. and Meister, K. (2004). Sampling design and sample selection through distribution theory. *J. Statist. Plann. Inference* **123**, 395-413.