Copyright © 1989 by the Genetics Society of America

Estimating the Variability of Substitution Rates

Michael Bulmer

Department of Statistics, Oxford University, Oxford OX1 3TG, England

Manuscript received March 9, 1989
Accepted for publication August 1, 1989

ABSTRACT

Suppose that amino acid or nucleotide data are available for a homologous gene in several species which diverged from a common ancestor at about the same time and that substitution rates between all pairs of species are calculated, correcting as necessary for multiple substitutions and for back and parallel substitutions. The variances and covariances of these corrected substitution rates are evaluated, and are used to construct a new test for uniformity (constancy of the molecular clock) and to find the best estimates of substitution rates in individual lineages with their standard errors. A substantial bias may arise if the effect of correcting the pairwise substitution rates is ignored.

THE variability of the molecular clock, that is to say of the rate of substitution at the amino acid or nucleotide level, is of interest both from the viewpoint of molecular evolution and in the reconstruction of phylogenetic trees from molecular data. KIMURA (1983) suggested that this question is most easily investigated from data on several species which all diverged from a common ancestor at nearly the same time, forming a so-called 'star phylogeny'; the radiation of the mammalian orders is generally supposed to be such a phylogeny, at least to a good approximation.

Suppose that sequence data are available for a homologous gene in s species which diverged from a common ancestor T years ago. Write λ_i for the substitution rate per year in the ith lineage, and $\alpha_i = \lambda_i T$ for the Expected number of substitutions per site. Direct estimates of the α_i's cannot be made, but the number of substitutions between each pair of species can be estimated, after correction if necessary for multiple substitutions and for back and parallel mutations. Thus the empirical data will be the $N = \frac{1}{2}s(s-1)$ estimates d_{ij} of the number of substitutions per site between species i and j $(i < j)$, so that

$$d_{ij} = \alpha_i + \alpha_j + e_{ij}. \qquad (1)$$

The sampling errors, e_{ij}, will have variances and covariances which must be known before the data can be analyzed. The variances, $\mathrm{Var}(d_{ij})$, are known from standard theory. For a star phylogeny the covariances with no lineage in common (such as $\mathrm{Cov}(d_{12},d_{34})$) are zero. The main technical problem is to evaluate the covariances with one lineage in common (such as $\mathrm{Cov}(d_{12}, d_{13})$). This will be done for amino acid and nucleotide substitutions in turn in the next section. [The covariances have previously been evaluated by NEI, STEPHENS and SAITOU (1985) and by NEI and JIN (1989) by a less formal argument.]

In the following section these calculations will then be used to obtain an improved method of analysing data on substitution rates. The variance/mean ratio (R) has become viewed as a critical test of the neutral theory of molecular evolution, values of R greater than unity indicating departures from this theory (GILLESPIE 1986). However, the usual formula for calculating R is substantially inflated because the process of correcting for multiple hits increases variability; the magnitude of this bias will be determined and a method of eliminating it presented. It will also be shown how to estimate within lineage substitution rates as accurately as possible, and how to calculate the standard errors of these estimates. Finally, a statistical test will be presented of the underlying assumption of a star phylogeny when data on four or more species are available.

COVARIANCES OF SUBSTITUTION RATES

Amino acid substitutions: Suppose that amino acid substitution follows a Poisson process with rate λ_i in the ith lineage, and that back mutation in the same lineage and parallel mutation in different lineages can be ignored. The probability that species i and j have the same amino acid at a particular site is then the probability that no mutation has occurred in either species since they split, which is equal to $\exp -(\alpha_i + \alpha_j)$. If p_{ij} is the observed proportion of amino acid substitutions between species i and j at n sites, then $d_{ij} = -\ln(1 - p_{ij})$ is an estimator of $(\alpha_i + \alpha_j)$.

To find the variances and covariances of the d_{ij}'s, define y_{ij} as a dummy variable which is 0 or 1 according as species i and j are identical (0) or differ (1) at a

The publication costs of this article were partly defrayed by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

particular site. Then

$$\begin{aligned} \text{Prob}(y_{ij}=0) &= \exp-(\alpha_i+\alpha_j), \quad i\neq j \\ \text{Prob}(y_{ij}=y_{ik}=0) &= \exp-(\alpha_i+\alpha_j+\alpha_k), \quad i\neq j\neq k \end{aligned} \tag{2}$$

whence

$$\begin{aligned} \text{Var}(y_{ij}) &= \exp-(\alpha_i+\alpha_j)(1-\exp-(\alpha_i+\alpha_j)) \\ \text{Cov}(y_{ij}, y_{ik}) &= E(y_{ij}y_{ik}) - E(y_{ij})E(y_{ik}) \\ &= \exp-(\alpha_i+\alpha_j+\alpha_k)(1-\exp-\alpha_i). \end{aligned} \tag{3}$$

Also, $p_{ij} = \sum y_{ij}/n$, the summation being over the n independent sites, so that

$$\begin{aligned} \text{Var}(p_{ij}) &= \text{Var}(y_{ij})/n \\ \text{Cov}(p_{ij}, p_{ik}) &= \text{Cov}(y_{ij}, y_{ik})/n. \end{aligned} \tag{4}$$

The approximate variances and covariances of the d_{ij}'s can be found by the standard Delta technique based on a Taylor series expansion for finding variances and covariances of functions of random variables (KENDALL and STUART 1963, Chapter 10). They are

$$\begin{aligned} \text{Var}(d_{ij}) &= (\exp(\alpha_i+\alpha_j)-1)/n \\ \text{Cov}(d_{ij}, d_{ik}) &= (\exp\alpha_i - 1)/n. \end{aligned} \tag{5}$$

Nucleotide substitutions: In estimating the nucleotide substitution rate it is necessary to allow for back and parallel substitutions as well as multiple substitutions. To obtain an analytic result I follow the model of JUKES and CANTOR (1969), supposing that nucleotide substitution occurs at rate λ_i in the ith lineage to one of the three other nucleotides, with the same rate of $\lambda_i/3$ to each of them at each site; a heuristic way of relaxing these rather restrictive assumptions will be described subsequently.

Let $P_{ij}(t)$ be the probability that species i and j differ at a particular site at time t, satisfying the differential equation

$$\begin{aligned} dP_{ij}/dt &= (1-P_{ij})(\lambda_i+\lambda_j) - \tfrac{1}{3}P_{ij}(\lambda_i+\lambda_j) \\ P_{ij}(0) &= 0 \end{aligned} \tag{6}$$

whose solution at time $t = T$ is

$$P_{ij} = \tfrac{3}{4}(1-\exp-\tfrac{4}{3}(\alpha_i+\alpha_j)). \tag{7}$$

P_{ij} can be estimated by p_{ij}, the observed proportion of differing sites, so that

$$d_{ij} = \tfrac{3}{4}\ln(1-\tfrac{4}{3}p_{ij}) \tag{8}$$

is an estimator of $(\alpha_i+\alpha_j)$. As before, p_{ij} is a binomial proportion with variance $P_{ij}(1-P_{ij})/n$; the approximate variance of d_{ij} based on the Delta technique (Taylor series expansion) is

$$\begin{aligned} \text{Var}(d_{ij}) = \tfrac{3}{4}\{&\tfrac{1}{4}\exp\tfrac{8}{3}(\alpha_i+\alpha_j) \\ &+ \tfrac{1}{2}\exp\tfrac{4}{3}(\alpha_i+\alpha_j) - \tfrac{3}{4}\}/n, \end{aligned} \tag{9}$$

as shown by KIMURA and OHTA (1972).

An argument set out in Appendix 1 shows that the covariance is

$$\begin{aligned} \text{Cov}(d_{ij}, d_{ik}) = \tfrac{3}{4}\{&\tfrac{1}{4}\exp\tfrac{8}{3}\alpha_i \\ &+ \tfrac{1}{2}\exp\tfrac{4}{3}\alpha_i - \tfrac{3}{4}\}/n. \end{aligned} \tag{10}$$

This equation can also be justified by the following informal argument. The covariance between d_{ij} and d_{ik} should depend only on events in the ith lineage and so should not depend on α_j or α_k. Equation 10 holds when $\alpha_j = \alpha_k = 0$, since in this case the covariance is the same as $\text{Var}(d_{ij})$, and should therefore hold for all values of α_j and α_k.

The Jukes-Cantor method assumes an equal substitution rate between different nucleotide pairs so that at equilibrium the four nucleotides are equally frequent. TAJIMA and NEI (1984) suggested that Equations 7 and 8 could be generalized to

$$\begin{aligned} P_{ij} &= b(1-\exp-(\alpha_i+\alpha_j)/b) \\ d_{ij} &= -b\ln(1-p_{ij}/b) \end{aligned} \tag{11}$$

where b allows for differences in nucleotide frequencies. Since p_{ij} is a binomial proportion, the approximate variance of d_{ij} as before is

$$\begin{aligned} \text{Var}(d_{ij}) = b\{&(1-b)\exp 2(\alpha_i+\alpha_j)/b \\ &+ (2b-1)\exp(\alpha_i+\alpha_j)/b - b\}/n, \end{aligned} \tag{12}$$

as obtained by TAJIMA and NEI (1984). By analogy with Equation 10 and the informal argument following that equation, it is conjectured that the covariance is given by

$$\begin{aligned} \text{Cov}(d_{ij}, d_{ik}) = b\{&(1-b)\exp 2\alpha_i/b \\ &+ (2b-1)\exp\alpha_i/b - b\}/n. \end{aligned} \tag{13}$$

The computer simulations of TAJIMA and NEI (1984) show that Equation 11 gives a reasonably good estimate for a wide range of substitution patterns. This question is discussed further by LEWONTIN (1989). Another complication arises from the need to estimate synonymous and nonsynonymous substitution rates separately, but Equations 12 and 13 can probably be used as adequate approximations with appropriate definition of n.

ANALYSIS OF DATA ON SUBSTITUTION RATES

Testing uniformity of substitution rates: Consider first the theoretical situation in which the number of substitutions between pairs of species can be observed directly with no need for correction. Let X_i be the total number of substitutions in the ith lineage which is assumed to be a Poisson variate with mean $n\alpha_i$, and write $x_i = X_i/n$. The observations are $d_{ij} = x_i + x_j$, with $\text{Var}(d_{ij}) = (\alpha_i+\alpha_j)/n$ and $\text{Cov}(d_{ij}, d_{ik}) = \alpha_i/n$. The

following identities hold between the x_i's and the d_{ij}'s:

$$x_i = (D_i - \tfrac{1}{2}s\bar{d})/(s-2) \quad (14a)$$

$$\bar{x} = \tfrac{1}{2}\bar{d} \quad (14b)$$

$$S_x = S_d/(s-2) \quad (14c)$$

where

$$D_i = \sum_{j>i} d_{ij} + \sum_{j<i} d_{ji} \quad (i \text{ fixed})$$

$$S_x = \sum_i (x_i - \bar{x})^2 \quad (15)$$

$$S_d = \sum_{i<j} (d_{ij} - \bar{d})^2.$$

If the substitution rates are the same in all lineages, $nS_x/\bar{x}$ is approximately a chi-square variate with $(s-1)$ degrees of freedom which can be used to test this hypothesis. From Equation 14, $2nS_d/(s-2)\bar{d}$ is an identical test statistic.

In practice d_{ij} will not be a direct measurement on $(x_i + x_j)$ but a corrected estimate of this quantity. The identities in Equation 14 cease to hold, but the expression on the right hand side of Equation 14a remains a sensible estimator of α_i, which we now denote a_i:

$$a_i = (D_i - \tfrac{1}{2}s\bar{d})/(s-2). \quad (16)$$

Either of the sums of squares S_a or S_d can be used as a basis for testing the null hypothesis, though they are not equivalent and their sampling distributions must be determined. It is shown in APPENDIX 2 that

$$(s-2)S_a/(v + (s-4)c) \quad (17)$$

is approximately a chi-square variate with $(s-1)$ degrees of freedom under the null hypothesis, where $v = \mathrm{Var}(d_{ij})$ and $c = \mathrm{Cov}(d_{ij}, d_{ik})$ as evaluated previously. (Note that with $\alpha_i = \alpha$ for all i, the variances have a common value, as do the covariances with one lineage in common. In evaluating v and c, α can be estimated by $\bar{a} = \tfrac{1}{2}\bar{d}$.) The distribution of S_d is more complicated and is not asymptotically chi-square, but its Expected value is

$$E(S_d) = \tfrac{1}{2}(s+1)(s-2)v - 2(s-2)c. \quad (18)$$

It is both simpler and more efficient to use S_a rather than S_d as a test statistic.

Equation 17 implies that the need to correct the observations for multiple (and possibly also for back and parallel) substitutions has inflated the expected value of S_a by the factor

$$\rho(S_a) = n(v + (s-4)c)/\alpha(s-2). \quad (19a)$$

The Expected value of S_d is inflated by the factor

$$\rho(S_d) = n(\tfrac{1}{2}(s+1)v - 2c)/\alpha(s-1). \quad (19b)$$

These inflation factors are evaluated in Table 1 for the Poisson formula appropriate to amino acid substitutions (taking values of v and c from Equation 5) and for the Jukes-Cantor formula appropriate for nucleotide substitutions (taking values of v and c from Equations 9 and 10). The inflation is substantial, and cannot be ignored, as suggested by KIMURA (1983, 1987), unless α is very small. It is likely to be particularly important in estimating synonymous substitution rates using the Jukes-Cantor formula. The synonymous substitution rate per site in a mammalian lineage since the origin of mammals is typically between 0.25 and 0.5, which gives an inflation factor of between two and fourfold for S_d.

TABLE 1

Inflation factors evaluated from Equation 19

	s				
α	3	4	5	6	
0.1	1.16	1.11	1.09	1.08	$\rho(S_a)$
0.25	1.46	1.30	1.24	1.22	Poisson formula
0.5	2.14	1.72	1.58	1.51	
1.0	4.67	3.20	2.70	2.46	
0.1	1.16	1.14	1.14	1.13	$\rho(S_d)$
0.25	1.46	1.40	1.38	1.36	Poisson formula
0.5	2.14	2.00	1.93	1.89	
1.0	4.67	4.18	3.93	3.78	
0.1	1.36	1.23	1.19	1.17	$\rho(S_a)$
0.25	2.21	1.76	1.61	1.53	Jukes-Cantor
0.5	5.36	3.56	2.96	2.66	formula
1.0	40.11	21.84	15.74	12.70	
0.1	1.36	1.32	1.30	1.28	$\rho(S_d)$
0.25	2.21	2.06	1.99	1.94	Jukes-Cantor
0.5	5.36	4.76	4.46	4.28	formula
1.0	40.11	34.02	30.97	29.15	

The values in Table 1 can also be interpreted as predicted values of R, the variance to mean ratio, calculated from the naive formulae $nS_a/(s-1)\bar{a}$ or $2nS_d/(s-1)(s-2)\bar{d}$. Thus values of R calculated in this way may well be as high as 4 for synonymous substitutions without invalidating the neutral theory. It is recommended that R should be calculated by dividing the formula in Equation 17 by $(s-1)$.

GILLESPIE (1986) developed an alternative method of correcting for multiple substitutions in amino acid data and of allowing for the effect of this correction on the variance. His method does not require the assumption of an underlying Poisson process, but the method developed here is both simpler and in other respects more general since it can be applied to nucleotide as well as amino acid data.

Estimating the substitution rates: If there is significant evidence of heterogeneity of the α_i's, then we will want to estimate these parameters as accurately as possible. The estimates a_i are unbiased estimators of the α_i's with variances and covariances with can be calculated as follows. First find the variances and covariances of the d_{ij}'s by substituting a_i for α_i in the

equations in the previous section; then find the variances and covariances of the a_i's from the standard formulas for linear functions of random variables:

$$\begin{aligned}\mathrm{Var}\left(\sum w_i y_i\right) &= \sum_i w_i^2 \mathrm{Var}(y_i) \\ &\quad + \sum_{i \neq j} w_i w_j \mathrm{Cov}(y_i, y_j) \\ \mathrm{Cov}\left(\sum w_i y_i, \sum u_i y_i\right) &= \sum_i w_i u_i \mathrm{Var}(y_i) \\ &\quad + \sum_{i \neq j} w_i u_j \mathrm{Cov}(y_i, y_j).\end{aligned} \tag{20}$$

The estimators a_i are unbiased and are likely to be reasonably efficient but they are not the minimum variance unbiased estimators since they give equal weight to all observations despite the fact that some observations are more accurate than others. Fully efficient estimators can be found by the method of weighted least squares. The model of Equation 1 can be written in matrix form as

$$\mathbf{d} = \mathbf{X}\alpha + \mathbf{e} \tag{21}$$

where $\mathbf{d}$ and $\mathbf{e}$ are the column vectors of observations and sampling errors respectively, α is the column vector of the α_i's and $\mathbf{X}$ is the $N \times s$ incidence matrix with two 1's in appropriate positions in each row and zero everywhere else. The method of weighted least squares finds the estimates $\hat{\alpha}$ which minimize the weighted sum of squares

$$SS = (\mathbf{d} - \mathbf{X}\alpha)^T \mathbf{V}^{-1} (\mathbf{d} - \mathbf{X}\alpha) \tag{22}$$

where $\mathbf{V}$ is the variance-covariance matrix of the d_{ij}'s. The estimates are

$$\hat{\alpha} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{d}, \tag{23}$$

with variance-covariance matrix

$$\mathrm{Var}(\hat{\alpha}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}. \tag{24}$$

If we substitute these estimates in Equation 22, the residual sum of squares is

$$SS = \mathbf{d}^T \mathbf{V}^{-1} \mathbf{d} - \alpha^T \mathbf{X}^T \mathbf{V}^{-1} \mathbf{d}. \tag{25}$$

This quantity has a chi-square distribution under normality with $N - s = \frac{1}{2}s(s - 3)$ degrees of freedom, which can be used to test the validity of some aspects of the model. If this statistic is significant, it can be concluded that either the assumption of a star phylogeny or the assumption of independent Poisson distributions is incorrect. No test is available unless there are more than three species, corresponding to the fact that an unrooted tree with three species can always be made into a star phylogeny by appropriate choice of the root.

LITERATURE CITED

GILLESPIE, J. H., 1986 Natural selection and the molecular clock. Mol. Biol. Evol. **3:** 138–155.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KENDALL, M. G., and A. STUART, 1963 *The Advanced Theory of Statistics*, Vol. 1. Charles Griffin, London.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KIMURA, M., 1987 Molecular evolutionary clock and the neutral theory. J. Mol. Evol. **26:** 24–33.

KIMURA, M., and T. OHTA, 1972 On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2:** 87–90.

LEWONTIN, R. C., 1989 Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6:** 15–32.

NEI, M., and L. JIN, 1989 Variances of the average numbers of nucleotide substitutions within and between populations.

NEI, M., J. C. STEPHENS and N. SAITOU, 1985 Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2:** 66–85.

SEARLE, S. R., 1971 *Linear Models.* Wiley, New York.

TAJIMA, F., and M. NEI, 1984 Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1:** 269–285.

Communicating editor: W.-H. LI

APPENDIX 1

Define $y_{ij} = 0$ or 1 according as species i and j have the same or a different nucleotide at a particular site. The state with respect to three species i, j and k is defined by the triplet (y_{ij} y_{ik} y_{jk}). Only five states are possible: (0 0 0), (1 1 0), (1 0 1), (0 1 1) and (1 1 1), which will be indexed by the integers 1 to 5. Write $\pi_i(t)$ for the probability that the system is in state i at time t ($i = 1, \ldots, 5$) and $\boldsymbol{\pi}$ for the vector of these probabilities. Under the Jukes-Cantor model it will satisfy the differential equation:

$$\begin{aligned} d\boldsymbol{\pi}/dt &= \mathbf{A}\boldsymbol{\pi} \\ \pi_1(0) &= 1, \quad \pi_i(0) = 0, \quad i > 1 \end{aligned} \tag{1.1}$$

$$\mathbf{A} = \begin{bmatrix} -(\lambda_i + \lambda_j + \lambda_k) & \lambda_i/3 & \lambda_j/3 & \lambda_k/3 & 0 \\ \lambda_i & -(\lambda_i/3 + \lambda_j + \lambda_k) & \lambda_k/3 & \lambda_j/3 & (\lambda_j + \lambda_k)/3 \\ \lambda_j & \lambda_k/3 & -(\lambda_i + \lambda_j/3 + \lambda_k) & \lambda_i/3 & (\lambda_i + \lambda_k)/3 \\ \lambda_k & \lambda_j/3 & \lambda_i/3 & -(\lambda_i + \lambda_j + \lambda_k/3) & (\lambda_i + \lambda_j)/3) \\ 0 & 2(\lambda_j + \lambda_k)/3 & 2(\lambda_i + \lambda_k)/3 & 2(\lambda_i + \lambda_j)/3 & -2(\lambda_i + \lambda_j + \lambda_k)/3 \end{bmatrix}$$

The eigenvalues of **A** are: $\mu_1 = -4(\lambda_i + \lambda_j + \lambda_k)/3$, $\mu_2 = -4(\lambda_i + \lambda_j)/3$, $\mu_3 = -4(\lambda_i + \lambda_k)/3$. $\mu_4 = -4(\lambda_j + \lambda_k)/3$, $\mu_5 = 0$. The solution of the differential equation at time $t = T$ is

$$\pi = 1/16 \begin{bmatrix} 6 & 3 & 3 & 3 & 1 \\ -6 & -3 & -3 & 9 & 3 \\ -6 & -3 & 9 & -3 & 3 \\ -6 & 9 & -3 & -3 & 3 \\ 12 & -6 & -6 & -6 & 6 \end{bmatrix} \begin{bmatrix} \exp \mu_1 T \\ \exp \mu_2 T \\ \exp \mu_3 T \\ \exp \mu_4 T \\ 1 \end{bmatrix} \tag{1.2}$$

We now compute

$$\begin{aligned} \mathrm{Cov}(y_{ij}, y_{ik}) &= E(y_{ij}y_{ik}) - E(y_{ij})E(y_{ik}) \\ &= \pi_2 + \pi_5 - P_{ij}P_{ik} \end{aligned} \tag{1.3}$$

Equation 10 is obtained by the Delta technique.

APPENDIX 2

Suppose that **y** is a multivariate normal random vector with zero mean and variance-covariance matrix **V**. Consider the distribution of the quadratic form

$$Q = \mathbf{y}^T\mathbf{A}\mathbf{y}. \tag{2.1}$$

SEARLE (1971, Chapter 2) shows that

$$E(Q) = \mathrm{tr}(\mathbf{AV}). \tag{2.2}$$

He also shows that kQ is a chi-square variate if and only if $k\mathbf{AV}$ is idempotent, having in this case degrees of freedom equal to the rank of **A**; it follows that

$$k = \mathrm{rank}(\mathbf{A})/\mathrm{tr}(\mathbf{AV}) \tag{2.3}$$

is the only constant that need be considered.

Consider first the distribution of

$$S_d = \mathbf{d}^T\mathbf{A}\mathbf{d} \tag{2.4}$$

$c = \mathrm{Cov}(d_{ij}, d_{ik})$ in off-diagonal positions representing pairs of values with one lineage in common, and zero elsewhere. It follows that

$$E(S_d) = \mathrm{tr}(\mathbf{AV}) = (N - 1)v - 2(s - 2)c. \tag{2.5}$$

Furthermore $k\mathbf{AV}$ is not idempotent, so that no multiple of S_d has a chi-square distribution.

Consider now the distribution of

$$S_a = \mathbf{a}^T\mathbf{A}\mathbf{a} \tag{2.6}$$

where **a** is the column vector of the a_i's and **A** is a square matrix of order s and rank $(s - 1)$ having $(s - 1)/s$ on the diagonal and $-1/s$ elsewhere. Under the null hypothesis **V** has

$$\begin{aligned} v_a = \mathrm{Var}(a_i) = \{&2s^2 - 7s + 6)v \\ &+ 2(s - 2)(s^2 - 5s + 5)c\}/2(s - 1)(s - 2)^2 \end{aligned} \tag{2.7}$$

on the diagonal, and

$$c_a = \mathrm{Cov}(a_i, a_j) = (c - 1/2v)/(s - 1)(s - 2) \tag{2.8}$$

elsewhere. Define

$$\delta = v_a - c_a = [v + (s - 4)c]/(s - 2). \tag{2.9}$$

Then

$$E(S_a) = \mathrm{tr}(\mathbf{AV}) = (s - 1)\delta. \tag{2.10}$$

Also, $\mathbf{AV}/\delta$ is idempotent, so that S_a/δ is a chi-square variate with $(s - 1)$ degrees of freedom.