ELSEVIER Computational Statistics & Data Analysis 23 (1997) 321–334

COMPUTATIONAL STATISTICS & DATA ANALYSIS

A measure for total variability in multivariate normal distribution

Seppo Mustonen

Department of Statistics, P.O. Box 54, 00014 University of Helsinki, Finland

Received November 1995; revised April 1996

Abstract

The total dispersion (sum of variances) and the generalized variance (determinant of the covariance matrix) do not meet certain essential requirements as measures of variability in the multivariate normal distribution. Therefore an alternative measure which is a generalization of the total dispersion is introduced and its formal and statistical properties studied. This measure is genuinely dependent on the covariances and it grows monotonously when new variables are included. The measure is strongly scale-dependent and not invariant even in orthogonal transformations, but it is shown that any invariant measure would violate the above-mentioned monotony requirement.

Keywords: Generalized variance; Measure of variability; Multivariate normal distribution; Total variation

1. Introduction

Common real-valued measures for variation of a p-dimensional normal distribution $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ are

$$\operatorname{tr}\Sigma = \lambda_1 + \lambda_2 + \cdots + \lambda_p \tag{1}$$

and

$$|\Sigma| = \lambda_1\lambda_2 \ldots \lambda_p \tag{2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of the covariance matrix Σ. The former one is often called a *total dispersion* (Seber, 1984) and used as a measure of

PII S0167-9473(96)00042-4

variation in principal components analysis, for example. The latter one is called a *generalized variance* and it has an information theoretic background because the entropy of the multivariate normal distribution is

$$\frac{1}{2}(p\log 2\pi + \log|\Sigma| + p). \tag{3}$$

Both measures have serious disadvantages which should be well-known but seldom discussed in statistical literature. See, however, Kowal (1971) and Johnson and Wichern (1988). The most serious handicap of measure (1) is that it observes neither covariances nor correlations. For example, if $p = 3$ and

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix},$$

where ρ is high, say $\rho = 0.999$, (1) gives the value 3 while a correct value for total variation should be only slightly over 1 since the two last variables simply echo the variation represented by the first variable alone.

The generalized variance (2) has the value $(1 + 2\rho)(1 - \rho)^2 = 2.998 \times 10^{-6} \approx 0$ which is absurd, too. In fact, if we have a system of $p - 1$ variables with a decent positive value of the generalized variance and add one more variable which is linearly dependent on the previous ones, the generalized variance collapses to 0 and the entropy (3) breaks down to $-\infty$.

To remove such defects, we have studied certain alternatives by starting from a specific idea presented in the next section. Our proposal for a measure of multivariate dispersion is

$$\mathrm{Mvar}(\Sigma) = \max \sum_{i=1}^{p} \sigma^2_{i,12\ldots i-1} \tag{4}$$

where $\sigma^2_{i,12\ldots i-1}$ is the residual variance of the ith variable when the previous ones are held constant and where the maximum is sought over all permutations of variables.

If we, for a moment, omit the max operation in (4), it is natural to assume that the total variation can be measured by taking the variance of the first variable as such, then taking the residual variance of the second variable after removing the variation explained by the first variable, then taking the residual variance of the third variable after removing the variation explained by the two first variables, etc. and, finally by summing these (residual) variances.

The crucial question is: What is the right order of variables in this summation since the value of the sum depends on the order of the variables? Our answer to this question is: Take the maximum value (4).

Before giving formal support for Mvar(Σ) let us study its behaviour in special cases. We see immediately that in the univariate case this measure is the same as the variance. Thus, the unit of measurement is the variation in N(0, 1).

When $p = 2$, we have

$$\mathrm{Mvar}(\Sigma) = \sigma_1^2 + (1 - \rho^2)\sigma_2^2,$$

where ρ is the correlation coefficient and the order of the variables must be selected so that $\sigma_1 \geq \sigma_2$. For example, in case $\sigma_1 = \sigma_2 = 1$ we obtain $\mathrm{Mvar}(\Sigma) = 2 - \rho^2$ and this gives 2 if $\rho = 0$ and 1 if $|\rho| = 1$.

Fig. 1 shows how different measures of total variability behave in the two-dimensional case as a function of the standard deviation of the second variable.

In the previous special case of 3 variables with a constant correlation $\rho = 0.999$ we have

$$\mathrm{Mvar}(\Sigma) = 3 - \rho^2(3 + \rho)/(1 + \rho) \approx 1.0035.$$

Please note that in this case the lowest possible value of ρ is $-\frac{1}{2}$. Thus, Mvar(Σ) is always finite.

In the p-dimensional distribution we have

$$\mathrm{Mvar}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 = \mathrm{tr}\,\Sigma$$

when the variables are uncorrelated and in general

$$\max(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2) \leq \mathrm{Mvar}(\Sigma) \leq \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2,$$

where the lower limit is attained when all correlations are ± 1. If all variances are equal to 1, Σ is a correlation matrix and

$$1 \leq \mathrm{Mvar}(\Sigma) \leq p.$$

2. Background of the measure

The idea of this paper arose from a desire to find a maximally parsimonious representation for a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ from $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Any such a vector can be generated by a linear transformation from $\boldsymbol{U} = (U_1, \ldots, U_p)^{\mathrm{T}}$ where U is $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ in the form

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{U} + \boldsymbol{\mu}, \tag{5}$$

where $\boldsymbol{C}$ is a $p \times p$-matrix and

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}. \tag{6}$$

Decomposition (6) can be selected in different ways. It becomes, however, unique if $\boldsymbol{C}$ is lower triangular. Then (6) is a Cholesky decomposition. By default (6) is

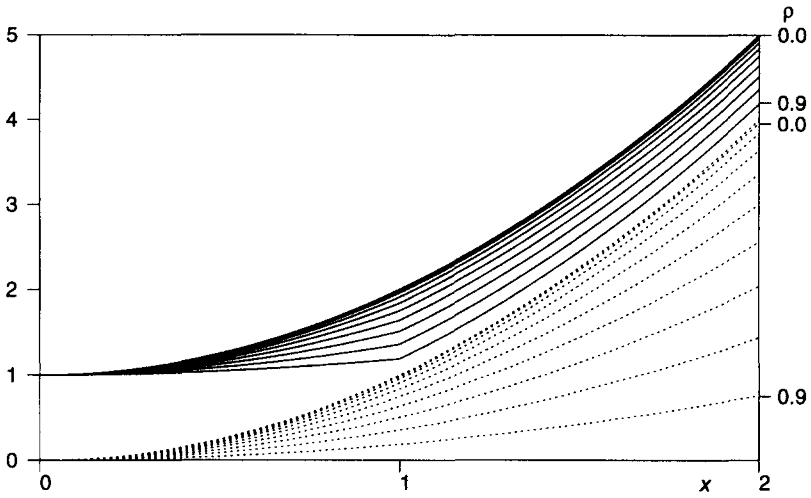


Fig. 1. Mvar(Σ) as a function of x for $\sigma_1 = 1$, $\sigma_2 = x$, $\rho = 0(0.1)\ 0.9$ (solid line), generalized variance for the same distributions (dotted line), Total dispersion (uppermost solid line).

unique for any non-singular $\boldsymbol{\Sigma}$. In singular cases where some of the pivot (diagonal) elements of $\boldsymbol{C}$ become 0, a unique representation is achieved by rearranging variables $X_1, \ldots, X_p$ so that when $r = \operatorname{rank}(\boldsymbol{\Sigma})$, the r first variables are linearly independent. Then the $p - r$ last columns in $\boldsymbol{C}$ are set $\boldsymbol{0}$.

To clarify the situation we write (5) componentwise as follows:

$$X_1 = c_{11}U_1 + \mu_1,$$

$$X_2 = c_{21}U_1 + c_{22}U_2 + \mu_2,$$

$$\ldots$$

$$X_p = c_{p1}U_1 + c_{p2}U_2 + \cdots + c_{pp}U_p + \mu_p.$$

It is natural to assume that this representation should be selected – by sorting X's – in such a way that each diagonal element of $\boldsymbol{C}$ 'dominates' elements below it, i.e.

$$c_{jj} \geq |c_{ij}|, \quad j = 1, \ldots, p-1, i = j+1, \ldots, p. \tag{7}$$

Then each U_j is introduced into the set of X variables with a 'great' weight c_{jj} and the later 'lesser' contributions of U_j may be neglected.

The maximal dominance, and parsimony, is attained when the X's are set in the order that maximizes

$$S = c_{11}^2 + \cdots + c_{pp}^2 \tag{8}$$

since the total dispersion (1) which has the same value for any order of variables can be written in the form

$$\operatorname{tr}\boldsymbol{\Sigma} = \operatorname{tr}\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} = \sum_{j=1}^{p}\sum_{i=j}^{p} c_{ij}^2 = \sum_{i=1}^{p} c_{ii}^2 + \sum_{i>j} c_{ij}^2.$$

Thus, maximization of (8) implies minimization of the sum of squares of the off-diagonal elements of $\boldsymbol{C}$.

On the other hand, the elements of $\boldsymbol{C}$ have simple statistical interpretations. In particular, we have

$$c_{ii} = \sigma_{i,12,\ldots,i-1}, \quad i = 1, \ldots, p \tag{9}$$

which means that maximization of (8) leads directly to measure $\mathrm{Mvar}(\boldsymbol{\Sigma})$.

Another important property of $\boldsymbol{C}$ is the following. If $\boldsymbol{C}$ is partitioned as

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \tag{10}$$

where $\boldsymbol{C}_{11}$ is $j \times j$ lower triangular and consequently $\boldsymbol{C}_{22}$ is $(p-j)\times(p-j)$ lower triangular, we get

$$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{C}_{22}\boldsymbol{C}_{22}^{\mathrm{T}} \tag{11}$$

where $\boldsymbol{\Sigma}_{22.1}$ is the partial covariance matrix of variables $X_{j+1}, \ldots, X_p$ when variables $X_1, \ldots, X_j$ are held constant.

To prove (9) and (11), we write (5) in a partitioned form

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(1)} \\ \boldsymbol{U}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \tag{12}$$

and study the conditional distribution of the $p-j$ component vector $\boldsymbol{X}^{(2)}$ when $\boldsymbol{X}^{(1)}$ is held constant, say, $\boldsymbol{X}^{(1)} = \boldsymbol{c}^{(1)}$.

We assume as before that if $r = \operatorname{rank}(\boldsymbol{\Sigma})$ is less than p, the variables are in such an order that the r first ones are linearly independent. By setting the last $p-r$ columns of $\boldsymbol{C}$ to $\boldsymbol{0}$, representation (12) becomes unique.

If now $\operatorname{rank}(\boldsymbol{\Sigma}_{11}) = \operatorname{rank}(\boldsymbol{C}_{11}) < j$, we have $\boldsymbol{C}_{22} = \boldsymbol{0}$. Thus in this case all variables of $\boldsymbol{X}^{(2)}$ are linearly dependent on $\boldsymbol{X}^{(1)}$ and (9) and (11) are trivially true. In the full rank case, $\operatorname{rank}(\boldsymbol{\Sigma}_{11}) = j$ and also $\boldsymbol{C}_{11}$ is non-singular. Hence, we obtain from (12)

$$\boldsymbol{X}^{(1)} = \boldsymbol{C}_{11}\boldsymbol{U}^{(1)} + \boldsymbol{\mu}^{(1)} = \boldsymbol{c}^{(1)}$$

which implies

$$\boldsymbol{U}^{(1)} = \boldsymbol{C}_{11}^{-1}(\boldsymbol{c}^{(1)} - \boldsymbol{\mu}^{(1)}).$$

Similarly, we have

$$\boldsymbol{X}^{(2)} = \boldsymbol{C}_{21}\boldsymbol{U}^{(1)} + \boldsymbol{C}_{22}\boldsymbol{U}^{(2)} + \boldsymbol{\mu}^{(2)} = \boldsymbol{C}_{22}\boldsymbol{U}^{(2)} + \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}(\boldsymbol{c}^{(1)} - \boldsymbol{\mu}^{(1)}) + \boldsymbol{\mu}^{(2)}.$$

This tells that in the conditional situation $\boldsymbol{X}^{(2)}$ is generated by a linear transformation from $p - j$ independent N(0, 1) variables $U_{j+1}, \dots, U_p$ with a coefficient matrix $\boldsymbol{C}_{22}$. Then the conditional distribution of $\boldsymbol{X}^{(2)}$ is multivariate normal with the covariance matrix $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{C}_{22}\boldsymbol{C}_{22}^{\mathrm{T}}$ which proves (11). Since $\boldsymbol{C}_{22}$ is lower triangular, (9) follows immediately.

3. Properties of the measure

In addition to features presented in the introduction, Mvar(Σ) satisfies the following essential condition. We use an alternative notation

$$\mathrm{Mvar}(\Sigma) = \mathrm{Mvar}(X_1, \dots, X_p)$$

where variables $X_1, \dots, X_p$ indicate the distribution in place of their covariance matrix Σ. Then

$$\mathrm{Mvar}(X_1, \dots, X_p) \geq \mathrm{Mvar}(X_1, \dots, X_{p-1}), \tag{13}$$

where the equality is possible only if X_p is linearly dependent on $X_1, \dots, X_{p-1}$. To prove this statement we represent X_p by 'linear regression' as

$$X_p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon,$$

where ε is $\mathrm{N}(0, \sigma^2)$ and independent of $X_1, \dots, X_{p-1}$. Then also $\sigma^2 = \sigma^2_{p,12\dots p-1}$ and

$$\mathrm{Mvar}(X_1, \dots, X_p) \geq \mathrm{Mvar}(X_1, \dots, X_{p-1}) + \sigma^2$$

which confirms (13). Since σ^2 is 0 only if X_p is linearly dependent on $X_1, \dots, X_{p-1}$, the equality is possible in this case only. However, even then the measure may grow if the variance of the new linearly dependent variable X_p is high enough.

The simplest example of such a situation is the following. Let X_1 be $\mathrm{N}(0, \sigma_1^2)$ and $X_2 = \alpha X_1$. Then $\mathrm{Mvar}(X_1) = \sigma_1^2$ and

$$\mathrm{Mvar}(X_1, X_2) = \sigma_1^2 = \mathrm{Mvar}(X_1) \quad \text{if } |\alpha| \leq 1$$

but

$$\mathrm{Mvar}(X_1, X_2) = \alpha^2\sigma_1^2 > \mathrm{Mvar}(X_1) \quad \text{if } |\alpha| > 1.$$

This feature is acceptable since no measure of variability can be scale-invariant. The important fact told by (13) is that Mvar(Σ) does not collapse like (2) when linear dependencies occur. Similarly, it is evident that

$$\mathrm{Mvar}(\alpha X_1, X_2, \ldots, X_p) \geq \mathrm{Mvar}(X_1, X_2, \ldots, X_p) \text{ when } |\alpha| > 1.$$

4. Computational aspects

Calculation of Mvar(Σ) means in principle an exhaustive search over all permutations of variables. One cannot assume that the magnitudes of variances in a straightforward way would determine the optimal order of variables. For example, the variable with the maximal variance is not necessarily the first one.

An exhaustive search over $p!$ permutations is infeasible in higher dimensions. Therefore, some short-cuts must be found. Fortunately, a simple stepwise procedure is available. It does not give the optimal solution in all cases but finds always a value which is close enough to Mvar(Σ).

The stepwise procedure is carried out p times by starting once from each variable. In round i, variable X_i is taken as the first one and the second variable is selected by maximizing

$$\sigma^2_{j,i}, j \neq i$$

with respect to j. The third variable is selected by maximizing

$$\sigma^2_{h,ij}, h \neq i, j$$

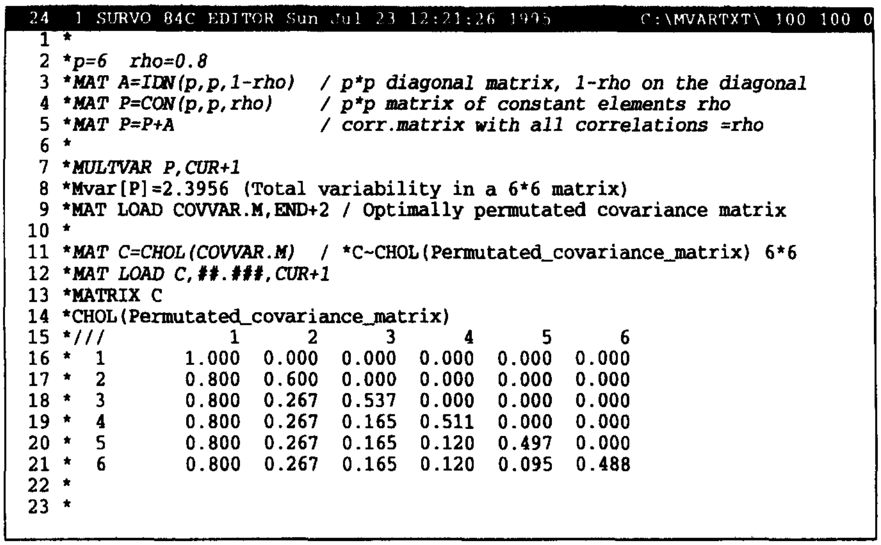
with respect to h, and so on. The solution is the one for which the sum of these residual variances is maximal.

In each of these p stepwise rounds, everything is obtained simply by performing the Cholesky decomposition stepwise. When selecting the kth variable, the residual variances to be compared are calculated by using (11). After selecting the kth variable, only $p - k$ last columns of the decomposition have to be updated.

A C program has been written for both the exhaustive and stepwise solution. The source code is freely available from the author. This code has been implemented in the SURVO 84C system (Mustonen, 1992) as an operation MULTVAR. By default this operation gives the stepwise solution. The stepwise solution is found in less than 2 seconds for $p = 20$ and in less than 5 minutes for $p = 60$ on a 486 PC (66 MHz). An exhaustive search already for $p \doteq 20$ would take almost 400 million years!

All computations in sequel have been performed by the MULTVAR operation. The simulation experiments have been carried out by making suitable macros on SURVO 84C.

The next display tells how Mvar(Σ) is computed:



```
24   1 SURVO 84C EDITOR Sun Jul 23 12:21:26 1995          C:\MVARTXT\ 100 100 0
 1 *
 2 *p=6  rho=0.8
 3 *MAT A=IDN(p,p,1-rho)  / p*p diagonal matrix, 1-rho on the diagonal
 4 *MAT P=CON(p,p,rho)    / p*p matrix of constant elements rho
 5 *MAT P=P+A             / corr.matrix with all correlations =rho
 6 *
 7 *MULTVAR P,CUR+1
 8 *Mvar[P]=2.3956 (Total variability in a 6*6 matrix)
 9 *MAT LOAD COVVAR.M,END+2 / Optimally permutated covariance matrix
10 *
11 *MAT C=CHOL(COVVAR.M)  / *C~CHOL(Permutated_covariance_matrix) 6*6
12 *MAT LOAD C,##.###,CUR+1
13 *MATRIX C
14 *CHOL(Permutated_covariance_matrix)
15 *///           1      2      3      4      5      6
16 *  1        1.000  0.000  0.000  0.000  0.000  0.000
17 *  2        0.800  0.600  0.000  0.000  0.000  0.000
18 *  3        0.800  0.267  0.537  0.000  0.000  0.000
19 *  4        0.800  0.267  0.165  0.511  0.000  0.000
20 *  5        0.800  0.267  0.165  0.120  0.497  0.000
21 *  6        0.800  0.267  0.165  0.120  0.095  0.488
22 *
23 *
```

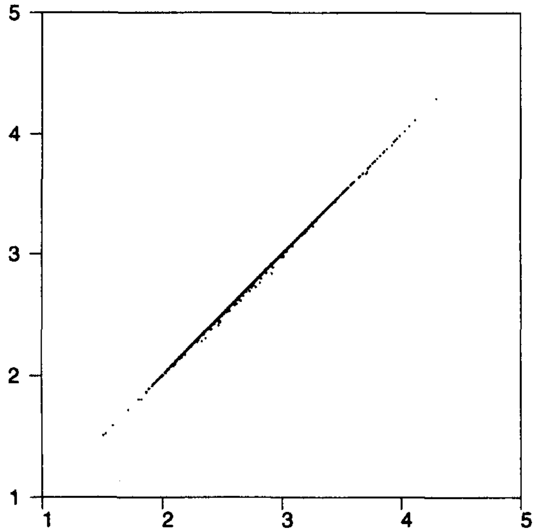


Fig. 2. Exhaustive vs. stepwise solution for 1000 random 7 × 7 covariance matrices.

All text and commands the user has typed in the edit field are shown here in *italics*. At first (lines 2–5) a 6 × 6 correlation matrix $\boldsymbol{P}$ with a constant correlation $\rho = 0.8$ is created by the matrix interpreter. The MULTVAR command on line 7 gives its results on lines 8–9. The optimal Cholesky decomposition is computed and printed by MAT commands on lines 11–12.

To check the efficiency of the stepwise solution, 1000 7×7 matrices A were generated. Each element of A's was a random number from a uniform distribution on $(-0.5, 0.5)$. For each A a 'covariance matrix' $\Sigma = AA^{\mathrm{T}}$ was computed and MULTVAR was applied to these Σ's by using both exhaustive and stepwise method.

Fig. 2 illustrates a good correspondence between the true solution and the stepwise solution. In 853 of those 1000 cases, the stepwise solution was optimal. The largest observed relative deviation was 2.6%.

Also other simulation experiments have been performed. All these trials show that relative error of the stepwise solution is insignificant in practice.

```
Stepwise_computation_of_Mvar()  /* in C pseudocode */
    {
    <<input: p x p covariance matrix S>>
    Mvar=-1.0; eps=0.000001;
    for (i=1; i<=p; ++i)
        {
        <<permutation vector q=(i,1,2,...,i-1,i+1,...,p)>>
        Mvar2=sum2_Cholesky(S,q);
        if (Mvar2>Mvar) { Mvar=Mvar2; <<optimal_q=q>> }
        }
    <<output: Mvar and optimal_q>>
    }
sum2_Cholesky(S,q)
    {
    sum2=S[q[1],q[1]];
    for (k=2; k<=m; ++k)
        {
        partial_Cholesky(C,S,q,k-1);
        if (C[q[k-1]][q[k-1]] <<not dominant in column k-1>>)
           return (-1.0);
        var_max=-1.0;
        for (h=k; h<=m; ++h) /* find max. res.variance */
            {
            s=0.0; /* see (11) */
            for (j=k to h) s=s+square_of(C[q[h]][q[j]]);
            if (s>var_max) { var_max=s; j_max=j; }
            }
        <<swap q[j_max] and q[k]>>
        sum2=sum2+var_max;
        }
    return(sum2);
    }
partial_Cholesky(C,S,q,h) /* update S=CC' from column h onwards */
    {
    for (i=h; i<=p; ++i) for (j=1; j<=i; ++j)
        {
        a=S[q[i]][q[j]];
        for (k=1; k<=j-1; ++k) a=a-C[q[i]][q[k]]*C[q[j]][q[k]];
        if (i==j)
            {
            b=sqrt(fabs(a)); if (b<eps) b=eps;
            C[q[i]][q[i]]=b;
            }
        else
            {
            b=C[q[j]][q[j]];
            if (b>0.0) C[q[i]][q[j]]=a/b;
                   else C[q[i]][q[j]]=0.0;
            C[q[j]][q[i]]=0.0;
            }
        }
    }
```

5. Examples and potential applications

As an example of how Mvar($\boldsymbol{\Sigma}$) behaves on different levels of dependency, we start from a $p \times p$ correlation matrix $\boldsymbol{P}$ with a constant correlation coefficient ρ for each pair of variables and let ρ vary between 0 and 1. In this case $\boldsymbol{C}$ in the Cholesky decomposition $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ is

$$\boldsymbol{C} = \begin{bmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & 0 & \dots & 0 \\ \beta_1 & \beta_2 & \alpha_3 & \dots & 0 \\ \beta_1 & \beta_2 & \beta_3 & \dots & 0 \\ . & . & . & \dots & . \\ \beta_1 & \beta_2 & \beta_3 & \dots & \alpha_p \end{bmatrix},$$

where

$$\alpha_i^2 = (1-\rho)[1+(i-1)\rho]/[1+(i-2)\rho], \quad i = 1, \dots, p$$

and

$$\beta_i^2 = \rho^2(1-\rho)/\{[1+(i-1)\rho][1+(i-2)\rho], i = 1, \dots, p-1.$$

The simple structure of C is detected from the numerical example in the previous chapter. The formulas for α_i and β_i are found by recursive computation and verified by induction. Then we have

$$\mathrm{Mvar}(\boldsymbol{P}) = (1-\rho) \sum_{i=1}^{p} [1+(i-1)\rho]/[1+(i-2)\rho] \tag{14}$$

since, due to complete symmetry, the sum of residual variances has a constant value. We can see from (14) that for large p and i values the contribution of a new variable tends to $1-\rho$.

This becomes evident also by noticing that a random vector $\boldsymbol{X}$ from N($\boldsymbol{0}, \boldsymbol{P}$) can be created from $p+1$ independent N(0, 1) variables $U_0, U_1, \dots, U_p$ by

$$X_1 = \sqrt{\rho}U_0 + \sqrt{1-\rho}\,U_1,$$
$$X_2 = \sqrt{\rho}U_0 + \sqrt{1-\rho}\,U_2,$$
$$\dots$$
$$X_p = \sqrt{\rho}U_0 + \sqrt{1-\rho}\,U_p.$$

Thus, there is a 'common factor' U_0 with 'communality' ρ and unique factors $U_1, \dots, U_p$ with variances $1-\rho$ each.

To give a better insight about the nature of Mvar($\boldsymbol{\Sigma}$), certain other special cases of the multivariate normal distribution deserve short comments. The following two results are conjectures based on extensive computational trials on the SURVO 84C system and seem to be valid at least for $p \le 90$.

If in (5) $\boldsymbol{C}$ is lower triangular with $C_{ij} = 1$ for all $i \geq j$ which means that X's are cumulative sums of independent N(0, 1) variables, the measure (4) has an interesting representation

$$\mathrm{Mvar}(\boldsymbol{\Sigma}) = p[\log_2 p/4 + 1 - \varepsilon(p)],$$

where $0 \leq \varepsilon(p) < 0.01$ and in particular $\varepsilon(p) \equiv 0$ when p is a power of 2. In this case (2) is 1 for all p and (1) is $p(p+1)/2$.

If a $p \times p$ correlation matrix $\boldsymbol{P}$ has the structure known from autoregressive models

$$\boldsymbol{P} = [\rho^{|i-j|}],$$

we have for all values of ρ an approximate linear representation

$$\mathrm{Mvar}(\boldsymbol{P}) \approx ap + b.$$

For small values of ρ, say $|\rho| < 0.5$, $a \approx 1 - \rho^2$ and $b \approx \rho^2$. In this case again the generalized variance (2) does not make sense since

$$\log |\boldsymbol{P}| = (p-1)\log(1-\rho^2)$$

(although linear) is a decreasing function of p.

Next we compare two covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ having the same correlation structure but different variances. When variation is measured by the generalized variance (2), Kowal (1971) pointed out that $\det(\boldsymbol{\Sigma}_1)/\det(\boldsymbol{\Sigma}_2)$ is a ratio of geometric means of the variances and does not depend on correlations at all. Kowal considered this an unfortunate property. The total dispersion (1) has the same disadvantage.

To indicate that $\mathrm{Mvar}(\boldsymbol{\Sigma})$ works better and observes correlations in a reasonable manner, we consider a special case with $p = 3$. Let us have

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \boldsymbol{\Sigma}_1 \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

By direct computation it is easy to show that $\mathrm{Mvar}(\boldsymbol{\Sigma}_2)/\mathrm{Mvar}(\boldsymbol{\Sigma}_1)$ is

$$[29 - 13\rho^2 - 4\rho^2(1-\rho)/(1+\rho)]/[3 - \rho^2(3+\rho)/(1+\rho)]. \tag{15}$$

Thus, (15) grows smoothly from 29/3 ($\rho = 0$) to 16 ($\rho = 1$). The ratio of generalized variances is 256 and the ratio of total dispersions 29/3 for all values $0 \leq \rho < 1$. The latter value coincides with (15) only for $\rho = 0$.

As a measure of variability $\mathrm{Mvar}(\boldsymbol{\Sigma})$ may also be useful in telling how much of variation of a given random vector is explained by another. Assume that $\boldsymbol{X}$ is $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partitioned into two subsets $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ where $\boldsymbol{X}^{(1)}$ is q-dimensional. Then

$$R^2(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = 1 - \mathrm{Mvar}(\boldsymbol{\Sigma}_{11.2})/\mathrm{Mvar}(\boldsymbol{\Sigma}_{11}) \tag{16}$$

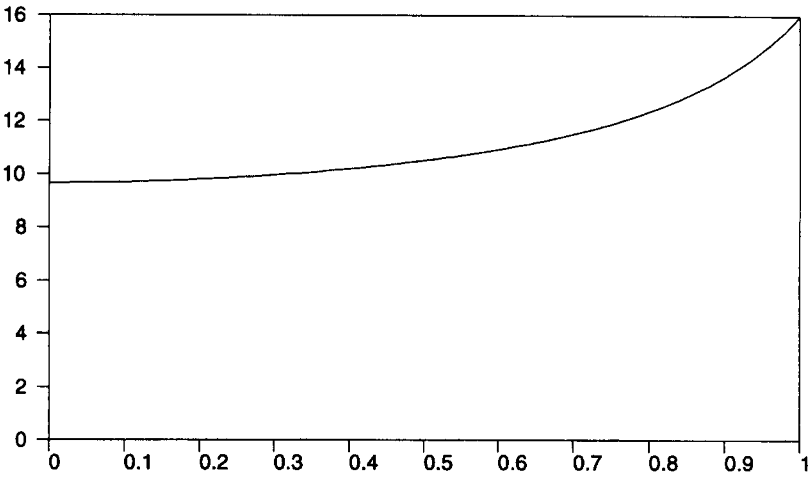


Fig. 3. Ratio $\mathrm{Mvar}(\Sigma_2)/\mathrm{Mvar}(\Sigma_1)$ as a function of ρ.

gives the proportion of the total variability of $X^{(1)}$ which is explained by $X^{(2)}$. In case $q = 1$ (16) is the same as the squared multiple correlation coefficient.

6. Estimation of Mvar(Σ)

In samples from $\mathrm{N}(\mu, \Sigma)$ an analogous estimate for $\mathrm{Mvar}(\Sigma)$ is $\mathrm{Mvar}(\boldsymbol{S})$ where $\boldsymbol{S}$ is the sample covariance matrix. To study properties of $\mathrm{Mvar}(\boldsymbol{S})$ a series of

Table 1

ρ	Mvar(**P**)	Mean of Mvar(**S**)	Relative bias	RMSE	Relative RMSE
0.0	5	4.908	– 0.018	0.328	·0.066
0.1	4.916	4.835	– 0.016	0.321	0.065
0.2	4.708	4.640	– 0.014	0.306	0.065
0.3	4.413	4.372	– 0.009	0.283	0.064
0.4	4.054	4.036	– 0.004	0.259	0.064
0.5	3.642	3.647	0.001	0.237	0.065
0.6	3.185	3.220	0.011	0.214	0.067
0.7	2.689	2.738	0.018	0.194	0.072
0.8	2.157	2.214	0.026	0.173	0.080
0.85	1.880	1.938	0.031	0.167	0.089
0.9	1.594	1.651	0.036	0.159	0.100
0.925	1.448	1.499	0.035	0.156	0.108
0.95	1.301	1.346	0.035	0.153	0.118
0.975	1.151	1.184	0.028	0.147	0.128
0.99	1.061	1.083	0.021	0.145	0.136
0.999	1.006	1.014	0.008	0.143	0.142

simulation experiments were performed in case $\boldsymbol{S} = \boldsymbol{P}$ with a constant correlation ρ and $p = 5$ for various values of ρ. The sample size was 100 and the experiment was repeated 10'000 times for each selected ρ value.

These experiments indicate that the distribution of Mvar($\boldsymbol{S}$) is close to normal for $\rho < 0.9$. For the greatest values of ρ it seems to be positively skewed. For the most values of ρ there is a slight bias. The main results are summarized in Table 1.

The bias is negative for ρ values up to about 0.5 and positive for greater ones as seen from Fig. 4.

In fact there must be a negative bias for $\rho = 0$ since with probability 1 Mvar($\boldsymbol{S}$) < tr($\boldsymbol{S}$) but tr($\boldsymbol{S}$) is an unbiased estimate of tr($\boldsymbol{P}$) = Mvar($\boldsymbol{P}$) for $\rho = 0$. When ρ tends to 1, the bias will tend to 0 since for $\rho = 1$, Mvar($\boldsymbol{S}$) = s^2 where s^2 is the common sample variance of variables. Fig. 5 tells how the relative RMSE grows with ρ.

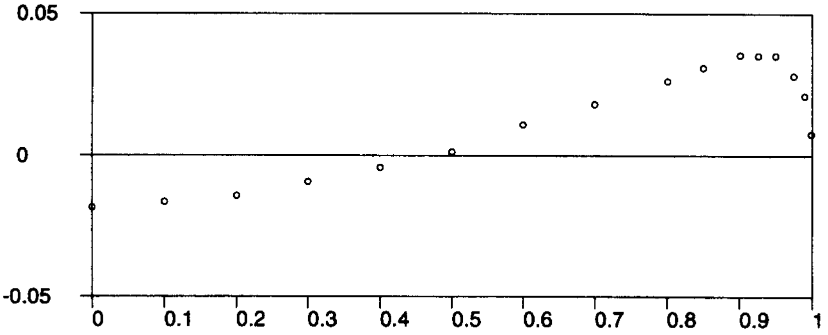


Fig. 4. Relative bias of Mvar($\boldsymbol{S}$) for various ρ values.

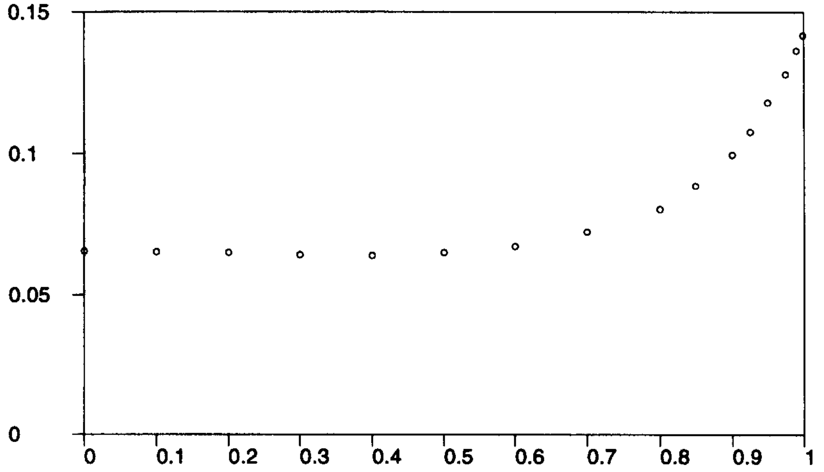


Fig. 5. Relative RMSE of Mvar($\boldsymbol{S}$) for various ρ values.

It is evident that Mvar($\boldsymbol{S}$) is a consistent estimator of Mvar($\boldsymbol{\Sigma}$). For example, if the sample size is 1000, for $\rho = 0.8$ the relative bias is only 0.012 and RMSE is 0.057. The dependency of the bias and RMSE on the sample size and $\boldsymbol{S}$ will be a target for a further study.

7. Concluding remarks

As pointed out earlier, Mvar($\boldsymbol{\Sigma}$) is strongly dependent on scales of variables. It is not even invariant in orthogonal transformations as measures (1) and (2) are. If such an invariance were required, it would mean that for a measure $M()$

$$M(\boldsymbol{\Sigma}) = M(\boldsymbol{T\Sigma T}^{\mathrm{T}})$$

for any orthogonal matrix $\boldsymbol{T}$. By selecting $\boldsymbol{T} = \boldsymbol{U}^{\mathrm{T}}$ from the spectral decomposition $\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U}^{\mathrm{T}}$ we would have $M(\boldsymbol{\Sigma}) = M(\boldsymbol{\Lambda})$ and the measure would be a function of eigenvalues as (1) and (2) are. Such a measure M based on eigenvalues only conflicts certain essential requirements.

For example, in case $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \boldsymbol{U\Lambda U}^{\mathrm{T}}$$

the eigenvalues are $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}$$

is the covariance matrix of principal components, say $\boldsymbol{Y} = (Y_1, Y_2)^{\mathrm{T}}$. If the measure M is invariant in an orthogonal transformation $\boldsymbol{Y} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$ we have $M(X_1, X_2) = M(Y_1, Y_2)$. In the special case $\rho = 1$ we should have $M(X_1, X_2) = M(X_1) = 1$ since $X_1 = X_2$ with probability 1 and in a univariate case, M gives the variance of the variable. Hence, we obtain an inequality

$$M(Y_1, Y_2) = M(X_1, X_2) = 1 < 2 = M(Y_1)$$

which violates the monotony requirement (13).

References

Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis*, 2nd edn. (Prentice-Hall, Englewood Cliffs, NJ, 1988).

Kowal, R.R., Disadvantages of the generalized variance as a measure of variability, *Biometrics*, **27** (1971) 213–216.

Mustonen, S., *SURVO, an integrated environment for statistical computing and related areas* (Survo Systems, Helsinki, 1992).

Seber, G.A.F., *Multivariate observations* (Wiley, New York, 1984).