ISSN 0494-7304 0132-053x

TARTU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО УНИВЕРСИТЕТА

ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

942

MULTIVARIATE STATISTICS

TARTU 1992

TARTU ÜLIKOOLI TOIMETISED
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
Alustatud 1893.a. VIHIK 942

MULTIVARIATE STATISTICS

Proceedings of the First Tartu Seminar, 27-29 August 1991

Tartu
1992

Tartu Ülikooli toimetised.
Vihik 942.
MULTIVARIATE STATISTICS.
Proceedings of the First Tartu Seminar
27-29 August 1991.
Tartu Ülikool.
EE2400 Tartu, Ülikooli 18.
Vastutav toimetaja T. Kollo.
10,12,9,5.T.295.400.
TÜ trükikoda. EE2400 Tartu, Tiigi 78.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 3-14.

ESTIMATES OF SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS BASED ON THE THEORY OF GENERAL STATISTICAL ANALYSIS

Alexander Babanin

Kiev University, Kiev, the UKRAINE

Keywords: *Limit theorems, random coefficients, regularized pseudosolution, G-condition*

1. INTRODUCTION

The solution of systems of linear algebraic equations (SLAE) is an important problem that arises in different scientific and engineering applications, such as numerical solution of differential and integral equations, planning of experiments, multivariate statistical analysis etc.

The number of original works in this field is very large. We point out the books by Wilkinson, (1965), Vojevodin, (1977), Tikhonov et al. (1990) which do not cover the variety of different methods of solution of SLAE and their applications. Notwithstanding the significant number of works and new achievements in the solution of SLAE of large dimension we can indicate the following problems:

1) the way of finding a consistent (the best in some sense) estimate of the SLAE solution is unknown if their coefficients are given with certain random errors;
2) the conditions of the existence of the moments of the components x_k of the vector x are not found;
3) under general conditions the limit theorems for the distribution functions of values x_k are not found.

Although a $(n \times m)$-matrix A is constant, it has to be treated as random by virtue of round-off errors of computer calculations (Girko (1990), p.286). We note that under dif-

ferent assumptions many limit theorems for the solutions of SLAE with random coefficients have been proved by Girko, (1980), (1988a), (1988b), (1989), (1990) and Girko, Babanin (1990). Among these limit theorems the most important is the so-called "arctangent law" (Girko (1990), p. 333).

Let the elements of a $(n \times n)$-matrix A and the components of a vector b be independent, their means be 0, their variances be 1, their absolute moments of order $4+\delta$, $\delta>0$ be bounded. Then

$$\lim_{n\to\infty} P\{x_k < z\} = \pi^{-1}\arctan z + 1/2,$$

where x_k are the components of the vector x (if $\det A=0$, then x_k are assumed to be equal to an arbitrary constant).

In the case, when the variances are bounded, the arctangent law was refined by Babanin, (1983). However, by means of these limit theorems the consistent estimates of the solutions of SLAE of the large order have not yet been obtained.

2. NEW ASSERTIONS FOR THE SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS WITH RANDOM COEFFICIENTS

Definition: *The sequence of the estimates $\hat{a}_{mn}$ of some value a_m is called G-consistent if*

$$\operatorname{plim}_{m,n\to\infty}[\hat{a}_{mn} - a_m] = 0.$$

2.1. The formulation of the problem. By means of independent observations X_i, $i=1,\dots,s$ under the matrix $A+C_2\Xi C_2$, $A=(a_{ij})$, $\Xi=(\xi_{ij})$, $i=1,\dots,n$, $j=1,\dots,m$ we have to find the G-consistent estimate of the regularized pseudosolution

$$d'X_\alpha = d'(C_1'C_1\alpha + A'C_2'^{-1}C_2^{-1}A\beta^{-1})^{-1}A'C_2'^{-1}C_2^{-1}b\beta^{-1}.$$

of the system of equations Ax=b, where C_1 and C_2 are $(n \times n)$- and $(m \times m)$-regular matrices respectively, $d \in R^m$, ξ_{ij}, $i=1,\dots,n$, $j=1,\dots,m$ are independent random elements for every value m and n, if the values σ_n^2, s_n, α_n, β_n depend on n and the G-condition holds:

$$\overline{\lim}_n \quad \sigma_n^2 n \beta_n^{-1} s_n^{-1} = t_1 < \infty,$$

$$\overline{\lim}_n \quad \sigma_n^2 m_n \beta_n^{-1} s_n^{-1} = t_2 < \infty,$$

$$\overline{\lim}_n \quad m_n n^{-1} = t_3 < \infty,$$

$$\sigma_n^2 = \mathrm{Var}\, \xi_{ij}.$$

2.2. Some remarks. Note that t_1 can be equal to 0. Later on we call the G-consistent estimates G-estimates. For simplicity index n at values σ_n, m_n, β_n, s_n is omitted.

If $\alpha=0$, the A'A matrix is regular, then

$$d'X_0 = d'(A'C_2'^{-1}C_2^{-1}A)^{-1}A'C_2'^{-1}C_2^{-1}b.$$

If $\alpha=0$ and A is a square matrix, then

$$d'X_0 = d'A^{-1}b.$$

Also, it is known that for some G-estimates G_n of values X_α

$$\mathrm{plim}_{n\to\infty} |X_\alpha - G_n|^2 = 0.$$

The parameter n is chosen artificially as the parameter of the limit transition in all transformations of G-estimates. It was necessary to use it for the proof of the existence of the G-estimates. We denote them $\tilde{G}_\beta$. The following $\tilde{G}_\beta$-estimate of G_β-class is found

$$\tilde{G}_\beta = \mathrm{Re}\, d'[C_1'C_1(\hat{\theta}+i\varepsilon) + \beta^{-1}(C_2^{-1}Z_\beta)'C_2^{-1}Z_\beta]^{-1}(C_2^{-1}Z_\beta)'C_2'^{-1}b\beta^{-1},$$

$$Z_\beta = \beta^{-1}\sum_{k=1}^{\beta} X_k.$$

θ is any real measurable solution of the equation

$$f_n(\theta) = \alpha, \tag{1}$$

$$f_n(\theta) = \theta \mathrm{Re}[1+\delta_1 s(\theta)]^2 - \alpha \mathrm{Im}[1+\delta_1 s(\theta)]^2 +$$
$$+ (\delta_1 - \delta_2)[1+\delta_1 \mathrm{Re}\, s(\theta)],$$

$$s(\theta) = n^{-1}\mathrm{Tr}[I(\hat{\theta}+i\alpha) + \beta^{-1}(C_2^{-1}Z_2C_1^{-1})'C_2^{-1}Z_2C_1^{-1}]^{-1},$$

$$\delta_1 = \sigma^2 n \beta^{-1} s^{-1},$$

$$\delta_2 = \sigma^2 m \beta^{-1} s^{-1}.$$

I is the identity matrix of the order m. In general the solution of equation (1) is non-unique. The $\tilde{G}_8$-estimate is rather distinct from the standard estimates of the form

$$d'X_{\mathrm{stand}} = d'[C'_1C_1\alpha + \beta^{-1}(C_2^{-1}Z_2)'C_2^{-1}Z_2]^{-1}(C_2^{-1}Z_2)'C_2^{-1}b\beta^{-1}.$$

The standard estimates have the biases which do not tend to zero as $n \to \infty$. These biases can be very significant.

3. NEW LIMIT THEOREMS FOR THE SOLUTIONS OF SLAE WITH RANDOM COEFFICIENTS

Theorem 1. *Let for each value n=1,2,... the elements* ξ_{pl}, $p=1,\dots,n$, $l=1,\dots,m$ *of a matrix* Ξ *be independent,* $E\xi_{pl}=0$, $\mathrm{Var}\xi_{pl}=\sigma^2$, *the G-condition holds,*

$$\lambda_m + \alpha \geq h,$$

where h is some nonnegative number,
$\lambda_1 \geq \dots \geq \lambda_m$ *are the eigenvalues of the matrix* $\tilde{A}'\tilde{A}\beta^{-1}$,

$$\tilde{A} = C_2^{-1} A C_1^{-1},$$

$$\overline{\lim}_{n\to\infty}[(\tilde{b}'\tilde{b} + \tilde{d}'\tilde{d})\beta^{-1/2} +$$

$$+ \sup_{k=1,\dots,n} a_k' a_k (\tilde{b}'\tilde{b}\tilde{d}'\tilde{d})^{1/2}\beta^{-1} < \infty,$$

where a_k *are the row vectors of the matrix* $\tilde{A}$, *and*

$$\tilde{b} = C_2^{-1} b,$$

$$\tilde{d} = C_1^{-1} d, \quad \sup_n \lambda_1 < \infty,$$

for some $\delta > 0$

$$\sup_l \sup_{p=1,\dots,n,\ l=1,\dots,m} E|\xi_{pl}|^{4+\delta} < \infty.$$

Then if $\epsilon \neq 0$

$$\mathrm{plim}_{n\to\infty}[\tilde{G}_n - \mathrm{Re}\, d'X_\epsilon + i\gamma(\epsilon)] = 0,$$

where

$$\gamma(\epsilon) = \epsilon\, \mathrm{Re}[1+\delta_1 s(\hat{\theta})]^2 + \hat{\theta}\, \mathrm{Im}[1+\delta_2 s(\hat{\theta})]^2$$

$$(\delta_1 - \delta_2)\delta_1 \mathrm{Im}\, s(\hat{\theta}).$$

Theorem 2. *If in addition to the conditions of Theorem 1*

$$\alpha + \lambda_m \geq 2\delta_2 + C,$$

$$2\delta_2(1 + \delta_2\tau)^2[\alpha + |\delta_1 - \delta_2|(1 + \delta_2\tau)] \times$$
$$\times (\alpha + \lambda_m - 2\delta_2)^2 + |\delta_1 - \delta_2|\delta_2\tau^2 \leq h < 1,$$

where $\tau = (\alpha + \lambda_m - \delta_2)^{-1}$, $C > 0$,

then $\lim_{\epsilon \to 0} \mathrm{plim}_{n \to \infty} |\gamma(\epsilon)| = 0$.

Corollary 1. *If the conditions of Theorem 2 hold, then*

$$\lim_{\epsilon \to 0} \mathrm{plim}_{n \to \infty} [G_\epsilon - d'X_n] = 0.$$

Corollary 2. *If the conditions of Theorem 1 hold,* $m = n$, $\lambda_m \geq 2\delta_2 + h$, $h > 0, \alpha = 0$, *then*

$$\lim_{\epsilon \to 0} \mathrm{plim}_{n \to \infty} [G_\epsilon - d'\tilde{A}^{-1}b] = 0.$$

4. NUMERICAL EXPERIMENTS

To compare the standard regularized estimates and Gs-estimates the algorithms were programmed in FORTRAN. The computations were carried out on a PC AT compatible computer. Some previous experiments were described in a paper by Girko, Babanin, (1989). These were carried out on a BESM-6 computer which is known to have excellent precision (10.E(-19)).

The following example is a net analog of an ill-posed problem from Fredholm (Tikhonov, et al., (1990)). As a test equation the following integral equation of the first kind

$$\int_f^g K(t,s)x(s)ds = h(t), \quad c \le t \le d \tag{2}$$

with the kernel

$$K(t,s) = \frac{1}{1+100(t-s)^2}$$

and the values f=0, g=1, c=-2, d=2 were considered. The exact solution for testing is given by equality (Tikhonov, et al., (1990), p. 34)

$$x(s) = (\exp\{-(s-0.3)^2/0.03 + \exp\{-(s-0.7)^2/0.03\})/0.9550408 - 0.0521309113 .$$

The values of the vector on the right hand side on the net $\{t_i\}$, i=1,....,m on the interval [c,d] were defined as the product of a $(n \times m)$-matrix A (approximating the operator in (2)) by the column vector x of the values of the exact solution on the net $\{s_i\}$, i=1,....,n on the interval [f,g]

$$b_i = \sum_{j=1}^{m} A_{ij}x(s_j).$$

$$\Phi(x) = |Ax - b|^2 .$$

This way of choosing the right hand side of the equality (2) guarantees that the minimum of the discrepancy functional on the corresponding set of vectors will vanish.

The matrix of the linear operator which is approximating the integral operator in equation (2) was chosen in the form

$$A_{ij} = \begin{cases} h_j K(x_i,s_j), & j=2,...,n-1; \\ \frac{1}{2}h_j K(x_i,s_j), & j=1,n; \end{cases}$$

where $h_s=(g-f)/(n-1)$ is a step of the uniform net $\{s_j\}$, $j=1,\ldots,n$ on $[f,g]$, $s_1=f$, $s_n=g$.

In Tikhonov et al. (1990), the accuracies of the right hand side and the operator in (2) were given by $\delta^2=10.E(-8)$ and $h^2=10.E(-10)$, respectively. It corresponds to the matrix perturbation of normally N(0, 3.E(-11)) distributed numbers.

In our experiment the observation matrix X is modelled in the following way: $X = A + \Xi$ where $\Xi = (\xi_{ij} n^{-1/2})$, $i=1,\ldots,n$, $j=1,\ldots,m$ is matrix of pseudorandom vectors. They are generated by means of a standard subroutine and normally N(0,3.E(-7)) distributed. The accuracy in the operator was 2.24E(-5), the accuracy on the right hand side was 3.14E(-4) and 6.41E(-6) as in Tikhonov, et al. (1990), p. 34.

First we obtained the standard regularized estimate using software from the book cited above. After that the double regularization procedure was carried out:

1) the starting point $\theta_0(\alpha)$ was the last value of the regularization parameter α, i.e. $\theta_0(\alpha)=\alpha$;
2) the method of successful approximations by the formula

$$\theta_k(\alpha) = \frac{\theta_{k-1}(\alpha)}{(1+a(\theta_{k-1}(\alpha))^2}, \quad k = 1,2,\ldots$$

was used, where $\theta(\alpha)$ is a solution of the equation

$$\theta(\alpha)[1 + \sigma^2 a(\theta(\alpha))]^2 + \sigma^2(1 - m/n)(1 + \sigma^2 a(\theta(\alpha))) = \alpha,$$

$\alpha > 0$, $a(\theta(\alpha)) = n^{-1}\mathrm{Tr}[I\theta(\alpha) + X'X]^{-1}$

(the Gs-estimate for the real case was considered);

3) the iterative process was terminated as soon as the condition

$$|\theta_k(\alpha) - \theta_{k-1}(\alpha)| \le 10^{-10}$$

was satisfied:

4) the desired value $\theta^*(\alpha)$ was substituted for the Gs-estimate

$$G_s = [\theta^*(\alpha) + X'X]^{-1}X'b.$$

The discrepancy for Gs-estimate is better (in 2.13 ...1133.01 times) than the discrepancy for the regularized estimate.

Table 1

Number of iterations	Parameter θ	Discrepancy Gs	dis stand dis Gs
0	α=θo 8.148E(-3)	dist st 4.167E(-4)	
1396	5.063E(-3)	1.949E(-4)	2.138
2265	3.189E(-3)	9.033E(-5)	4.613
3000	1.651E(-3)	2.877E(-5)	14.483
3330	9.815E(-4)	1.132E(-5)	36.813
3500	6.448E(-4)	5.310E(-5)	78.472
3700	2.597E(-4)	1.169E(-6)	356.615
3746	1.738E(-4)	6.936E(-7)	600.807
3800	7.557E(-5)	3.678E(-7)	1133.014

In Figure 1 representing Table 2 the exact solution is shown by asterisks. the standard (approximate) solution is shown by hats. the Gs-solution is shown by plusses.

Table 2

EXACT SOLUTION=

.0000022	.0320500	.0782515	.1415692	.2238934	.3251580
.4425374	.5699922	.6984153	.8165286	.9124924	.9759422
1.0000450	.9830439	.9289117	.8469308	.7502968	.6540659
.572875	.5188388	.4999051	.5188388	.5728755	.6540659
.7502986	.8469308	.9289117	.9830439	1.0000450	.9759422
.9124924	.8165296	.6984153	.5699922	.4425374	.3251580
.2238934	.1415692	.0782515	.0320500	.0000022	

STANDARD SOLUTION AS SIGMA= .57730202979709D-03

.1006648	.1341703	.1795270	.2249798	.2843824	.3486916
.4300577	.5084717	.5875654	.6398337	.6842951	.7316810
.7496157	.7540043	.7280539	.7046086	.6847064	.6518280
.6184477	.5977486	.5994236	.6019248	.6202396	.6468041
.6850821	.7044790	.7261115	.7462028	.7523570	.7261787
.6837627	.6398396	.5824952	.5111414	.4255976	.3566069
.2832791	.2255258	.1766372	.1376967	.1033248	

SOLUTION G8 =

.0201848	.0541344	.1068443	.1635703	.2407649	.3322013
.4542393	.5723326	.6920770	.7786502	.8481158	.9172177
.9378989	.9280398	.8654959	.8055157	.7439806	.6725418
.6047111	.5565607	.5499901	.5646754	.6111112	.6717731
.7481335	.8050075	.8603216	.9129163	.9342202	.9031792
.8444219	.7752701	.6860345	.5803188	.4525894	.3462920
.2437124	.1645387	.1029640	.0583629	.0193878	

ERRORS: - IN OPERATOR: .22400000000000D-04
IN RIGHT HAND SIDE: .31400000000000D-03
STOP CODE: 0
DISCREPANCY: .416709573D-03
REGULARIZATION PARAMETER: .814805728D-02
PARAMETER TETA: .124305508D-02
DISCREPANCY G8: .173291032D-04
NUMBER OF ITERATIONS: 3200
RELATION AN2/AN2G8: .240468054D+02
ACCURACY: .198856735D-09

Figure 1

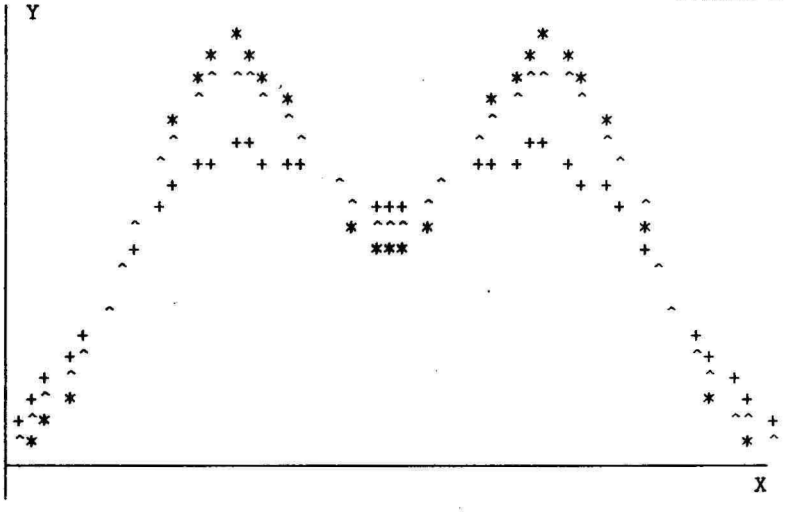

Table 3

σ=5.773E(-4) · σ^2=3.E(-7)

h^2=2.24E(-5) σ^2=6.41E(-6)

Number of iterations	Parameter θ	Discrepancy Gs	dis stand dis Gs
0	$\alpha=\theta_o$ 1.117E(-3)	dist st 3.084E(-5)	
500	7.098E(-4)	6.303E(-6)	4.893
600	5.143E(-4)	3.556E(-6)	8.674
700	3.223E(-4)	1.622E(-6)	19.010
800	1.354E(-4)	5.383E(-7)	57.298
815	1.080E(-4)	4.498E(-7)	68.566
825	9.002E(-5)	4.013E(-7)	76.858

References

Babanin, A.S. (1983). A generalization of arctangent law. Theor. Prob and Math. Stat., **28**, 3-5.

Girko, V.L. (1980). An arctangent law. Reports of the Academy of Sciences of the Ukrainian SSR. **A**, **4**, 7-9. (in Ukrainian)

Girko, V.L. (1984). A circle law. Theor. Prob. and Appl. **29**, 669-679.

Girko, V.L. (1988). Multivariate Statistical Analysis. Vyshcha Shkola, Kiev (in Russian).

Girko, V.L. (1990). Theory of Random Determinants. Kluwer, Amsterdam.

Girko, V.L., Babanin A.S. (1988). On the G_B-estimate of the solutions of empirical systems of linear algebraic equations. Theor. Prob. and Math. Stat., **39**, 29-33.

Tikhonov A.N., Goncharsky A.V., Stepanov V.V. and Yagola A.G. (1990). Numerical Methods for the Solution of Ill-Posed Problems. Nauka, Moscow (in Russian).

Voievodin V.V. (1977). Computational Bases of Linear Algebra. Nauka, Moscow (in Russian).

Wilkinson I.N. (1965). The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 15-20.

COMPUTING THE DISTRIBUTION OF A QUADRATIC FORM IN NORMAL VARIABLES: A SURVEY OF RECENT DEVELOPMENTS

Richard William Farebrother

University of Manchester, Manchester, ENGLAND

Keywords: *Characteristic function, Durbin-Watson statistic, linear combination of χ^2 random variables*

1. INTRODUCTION

A wide class of statistical problems directly or indirectly involve the evaluation of probabilistic expressions of the form.

$$\Pr(u'Pu < x) \qquad (1.1)$$

where u is an nx1 matrix of random variables which is normally distributed with mean θ and variance Ω, and where x is a scalar, θ is an nx1 matrix, P is an nxn symmetric matrix, and Ω is an nxn symmetric positive definite matrix.

In this paper we will outline recent developments concerning the "exact" evaluation of this expression. In section 2 we will discuss implementations of standard procedures which involve the diagonalisation of the nxn matrix P whilst in section 3 we will discuss procedures which do not require this preliminary diagonalisation. Finally in section 4 we will examine procedures which may be used to evaluate the joint distribution of several quadratic forms in the same set of normal variables.

2. DIAGONAL QUADRATIC FORMS

Let L be an nxn lower triangle matrix such that $\Omega = LL'$, let $Q = L'PL$ and let $v = L^{-1}u$ then $u'Pu = v'Qv$ and we have to evaluate

$$\Pr(v'Qv < x) \qquad (2.1)$$

where $v = L^{-1}u$ is normally distributed with mean $\mu = L^{-1}\varsigma$ and variance $I_n = L^{-1}\Omega\,(L')^{-1}$.

Now let H be an nxn orthonormal matrix such that $T = H'QH$ is tridiagonal and let $w = H'v$ then $v'Qv = w'Tw$ and we have to evaluate

$$\Pr(w'Tw < x) \qquad (2.2)$$

where $w = H'v$ is normally distributed with mean $\kappa = H'\mu$ and variance $I_n = H'H$.

Finally, let G be an nxn orthonormal matrix suchthat $D = G'TG$ is diagonal and let $z = G'w$ then $w'Tw = z'Dz$ and we have to evaluate

$$\Pr(z'Dz < x) \qquad (2.3)$$

where $z = G'w$ is normally distributed with mean $\upsilon = G'\kappa$ and variance $I_n = G'G$.

Thus expressions (1.1), (2.1) or (2.2) may be evaluated indirectly by evaluating

$$\Pr\left[\sum_{j=1}^{n} d_{jj}\, z_j^2 < x\right] \qquad (2.4)$$

Now the characteristic function of the weighted sum of noncentral $\chi^2(1)$ variables $z'Dz$ is given by

$$\phi(t) = [\psi(f)]^{-\frac{1}{2}} \qquad (2.5)$$

where $f = 2it$ and

$$\psi(f) = \det(I - fD)\exp\left[\nu'\nu - \nu'(I - fD)^{-1}\nu\right] \qquad (2.6)$$

Thus expressions of the form (1.1) may be evaluated by applying the Imhof (1961), Ruben (1962) or Grad and Solomon (1955) procedures to equation (2.4).

(a) The standard Imhof procedure is a general procedure which obtains the desired result by numerical integration. It has been programmed in Fortran by Koerts and Abrahams (1969) and in Algol and Pascal by Farebrother (1990). An improved version of the Imhof procedure has been programmed in Algol by Davies (1980). A Fortran translation is

available from R B Davies and Pascal translations from J C Nankervis of the Central London Polytechnic and the present author.

(b) The Ruben procedure expands expression (2.4) as a sum of central χ^2 distribution functions but is restricted to positive definite matrices. It has been programmed in Fortran by Sheil and O'Muircheartaigh (1977) and in Algol and Pascal by Farebrother (1984a).

(c) The Grad and Solomon procedure (sometimes also attributed to Pan) uses contour integration to evaluate expression (2.5) but is restricted to sums of central $\chi^2(1)$ variables with distinct weights. It has been programmed in Algol by Farebrother (1980, 1984b). A Fortran translation is available from M J Harrison of Trinity College, Dublin, and a Pascal translation from the author.

(d) The Ruben and Grad and Solomon procedures are very much faster than the Imhof procedure but not of such general application, see Farebrother (1984c) for details.

3. NONDIAGONAL QUADRATIC FORMS

The fundamental problem with the standard procedures outlined insection 2 is that the matrices P,Q, and T have to be reduced to diagonal form before these methods can be applied. Recent advances in this area by Palm and Sneck (1984), Farebrother (1985), Shively, Ansley and Kohn (SAK, 1990) and Ainsley Kohn and Shively (AKS, 1991) are based on the observation that the transformed characteristic function $\psi(f)$ may also be written as

$$\psi(f) = \det(I-fT)\exp[\kappa'\kappa - \kappa'(I-fT)^{-1}\kappa] \quad (3.1)$$

$$\psi(f) = \det(I-fQ)\exp[\mu'\mu - \mu'(I-fQ)^{-1}\mu] \quad (3.2)$$

or

$$\psi(f) = \det(\Omega)\det(\Omega^{-1}-fP)\exp[\varsigma'\Omega^{-1}\varsigma - \varsigma'(\Omega^{-1}-fP)^{-1}\varsigma] \quad (3.3)$$

So that the numerical integration of the Imhof procedure may be performed using complex arithmetic.

(e) Farebrother's (1985) variant of Palm and Sneek's procedure is of

general application but it requires that Q be constructed and then reduced to tridiagonal form. It has been programmed in Pascal by Farebrother (1990).

(f) The SAK and AKS procedures are of more restricted application. It is assumed that P and Ω may be expressed as $P = BAB'$ and $\Omega = B\Sigma B'$ where A and Σ (or their inverses) are mxm symmetric band matrices and where B is an nxm matrix of rank n which satisfies $BC=0$ and $BB' = I_n$ for some given m x (m-n) matrix C of rank m - n.

In this context with $\delta=0$ and $x=0$, and with further restrictions on the form of Ω, SAK (1990) and AKS (1991) respectively used the modified Kalman filter and the Cholesky decomposition to evaluate expression (3.3) and thus (1.1) without forming P, Ω and Q and without reducing Q to tridiagonal form.

Both of these procedures are very much faster than the Davies (1980) and Farebrother (1990) procedures for large values of n but Farebrother (1991) has expressed reservations concerning their numerical accuracy as the matrix techniques they employ are known to be numerically unstable in certain circumstances.

The implementation of both procedures is specific to the particular class of A and Σ matrices selected, but KAS (1991) have programmed the AKS procedure in Fortran for the generalised Durbin-Watson statistic.

(g) See AKS (1991, Appendix A) for an excellent discussion of methods for choosing the correct square not in expression (2.5). This is an important practical problem which is too often ignored.

4. MULTIPLE QUADRATIC FORMS

Finally we note that Shephard (1991) and Shively (1989) have extended the standard numerical inversion procedure to multivariate problems of the form

$$\Pr\left[\bigcap_{j=1}^{k}(u'P_j u < x_j)\right] \tag{4.1}$$

Shively (1988) and Shephard (1990) have illustrated this procedure by using it to evaluate the probabilities associated with a test for the stability of regression coefficients and those associated with a test

for the equality of regression coefficients in submodels with distinct error variances.

REFERENCES

Ansley, C F, Kohn, R and T S Shively (1991), Computing p-values for the Durbin-Watson and other invariant test statistics, University of New South Wales, Discussion Paper.

Davies, R B (1980), The distribution of a linear combination of χ^2 random variables *Applied Statistics 29*: 323-333.

Farebrother, R W (1980), Pan's procedure for the tail probabilities of the Durbin-Watson statistic *Applied Statistics 29*: 224-227 and *30*: 189.

Farebrother, R W (1984a), The distribution of a positive linear combination of χ^2 random variables, *Applied Statistics 32*, 322-339.

Farebrother, R W (1984b), The distribution of a linear combination of central χ^2 random variables *Applied Statistics 33*: 363-366.

Farebrother, R W (1984c), The distribution of a linear conbination of χ^2 random variables, *Applied Statisitics 33*: 366-369.

Farebrother, R W (1985), Eigenvalue-free methods for computing the distribution of a quadratic form in normal variables, *Statitische Hefte 26*: 287-302.

Farebrother, R W (1990), The distribution of a quadratic form in normal variables *Applied Statistics 39*: 294-309

Farebrother R W (1991), "A critique of recent methods for computing the distribution of the Durbin-Watson and other invariant test statistics", University of Manchester, Discussion Paper.

Grad, A and Solomon, H (1955), Distributions of quadratic forms and some applications, *Annals of Mathematical Statistics* 26: 464-477.

Imhof J P (1961), Computing the distribution of a quadratic form in normal variables, *Biometrika 48*: 419-426

Koerts J and Abrahamse A P J (1969), *On the Theory and Application of the General Linear Model*, Rotterdam University Press, Rotterdam.

Kohn R, Ansley C F and Shively T S (1991), Computing p-values for the

generalized Durbin-Watson statistic and residual autocorrelations in regression, University of New South Wales, Discussion Paper.

Palm F C and Sneek J M (1984), Significance tests and spurious correlation in regression models with autorcorrelated errors, *Statistische Hefte 25*: 87-105

Ruben, H (1962) Probability content of regions under spherical normal distributions, IV: The distributions of homogeneous and nonhomogeneous quadratic functionsof normal variables, *Annals of Mathematical Statistics 33*: 542-570

Sheil J and O'Muircheartaigh, I (1977), The distribution of non-negative quadratic forms in normal variables, *Applied Statistics 26*: 92-98

Shephard N G (1990), Tabulation of Farebrother's test of linear regressions in linear regression models under heteroscedasticity, London School of Economics, Discussion Paper.

Shephard N G (1991), From characteristic function to distribution function: a simple framework for the theory, *Econometric Theory*, forthcoming.

Shively T S (1988) An analysis for tests for regression coefficient stability, *Journal of Econometrics 39*: 367-386.

Shively T S (1989) Numerical inversion of a bivarate characteristic function, University of Texas, Discussion Paper.

Shively T S, Ansley C F and Kohn R (1990), Fast evaluation of the Durbin-Watson and other invariant test statistics in time series regression *Journal of the American Statistical Association 85*: 676-685.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 21-27.

OPTIMAL k-CENTRES FOR A TWO-DIMENSIONAL NORMAL DISTRIBUTION

Sven Kipper and **Kalev Pärna**
University of Tartu, Tartu, ESTONIA

Keywords: *Discrete approximation, two-dimensional normal distribution, k-centers, Lloyd's algorithm*

1. Introduction. The aim of this paper is to present the results of the calculation of k-centres for a two-dimensional normal distribution. Here the term 'k-centre is used to denote the best k-point discrete approximation to a given probability distribution - a two-dimensional normal in our case. We shall show what are the best systems of $k = 1,2,\ldots,8$ points in R^2 that represent the given normal distribution as well as possible. A quadratic loss-function has been used to measure the distance between the normal distribution and an approximating system of k points.

The k-centres of the normal distribution (or of some other distribution of interest) can be used in the optimal allocation problem where the resources have to be placed at k points which have to be chosen in the best way to satisfy the normally distributed demand. Alternatively, k-centres are necessary in the information transmission when the channel is capable of admitting only k distinct values of the (continuous) signal. Then, again, these k values should be selected in an optimal way.

This paper is concerned with a very specific normal

distribution. We shall fix the mean value of the normally distributed vector (X,Y) at zero and take the standard deviations of the two independent components equal to σ_x=1.5 and σ_y=1. The k-centres which we obtain in this case can be transformed in an obvious manner to obtain k-centres for any other normal distribution having the same ratio of variances. But, unfortunately, our results can not be used for easy calculation of k-means for the normal distributions with the different ratio of variances. The reason why we cannot do so is, in geometrical terms, that the angles are not invariant under arbitrary linear transformations. However, due to the continuity of k-centres with respect to the initial distribution (see e.g. Pärna (1987), proposition 12.2; Cuesta and Matràn (1988)), the k-centres obtained here can be used, at least, as initial values in the iterative process for finding k-centres of the different but close distributions.

Note that the case when both variances, σ_x^2 and σ_y^2, are equal to 1 has been studied in Book (1991).

2. The k-centres. For the precise description of the problem it is necessary to introduce some notation. We shall consider a random vector Z = (X,Y) consisting of two independent, normally distributed components having zero means and standard deviations σ_x = 1.5 and σ_y = 1. Let us write $N_2(0,\Sigma)$ for this distribution. The corresponding density is

$$f(x,y) = (3\pi)^{-1}\exp\{-0.5(x^2/2.25 + y^2)\}. \qquad (1)$$

Let $A = \{a_1,\dots,a_k\}$ be a subset of R^2. We define its goodness by the **loss-function**

$$W(A) = \int\int \min_{1\le i\le k} [(x-a_{ix})^2 + (y-a_{iy})^2]\, f(x,y)\, dxdy, \qquad (2)$$

where a_{ix} and a_{iy} are the coordinates of a_i, and the domain of integration is the whole plane. The formula (2) can be rewritten as

$$W(A) = \int \min_{1\le i\le n} \| z-a_i \|^2 f(z)\, dz,$$

where z stands for the pair (x,y), or as

$$W(A) = E \min_{i \le k \le n} \| Z - a_i \|^2$$

- the expectation of the squared distance from the random point Z to its nearest point in A. The problem is to minimize W(A) over all possible choices of A keeping k fixed. Any optimal A will be called a '**k-centre**' for $N(0,\Sigma)$. The existence of at least one k-centre for a wide class of probability measures has been proved in Pärna (1990)). Note that this class includes the normal distribution, too.

Let W_k^* stand for the minimum value of W(A) over all k-sets. It is not difficult to prove that, for any distributions not concentrated at less than k points, the strict inequalities $W_1^* > W_2^* > \ldots > W_k^* > \ldots$ take place (see e.g. Pärna (1990)).

Further, for any k point set A, let us define its **Voronoi partition** $S(A) = \{S_1(A), S_2(A), \ldots, S_k(A)\}$, where

$$S_i(A) \subseteq \{z \in R^2 : \|z-a_i\| \le \|z-a_j\| \text{ for each } j,\ j \ne i\}, \qquad (3)$$

$$S_i(A) \cap S_j(A) = \emptyset, \quad \bigcup_{i=1}^{k} S_i(A) = R^2.$$

The Voronoi partition (known also as the minimum distance partition determined by A) allows us to express W(A) as a sum of integrals:

$$W(A) = \sum_{i=1}^{k} \int_{S_i(A)} \|z-a_i\|^2 f(z)dz, \qquad (4)$$

Moreover, it is easy to show that if A is a k-centre, the conditional expectation of Z in the Voronoi region $S_i(A)$ coincides with a_i itself. In formulas, it means that, for optimal A,

$$a_{ix} = \frac{\iint_{S_i(A)} x\, f(x,y)dxdy}{\iint_{S_i(A)} f(x,y)dxdy}, \qquad a_{iy} = \frac{\iint_{S_i(A)} y\, f(x,y)dxdy}{\iint_{S_i(A)} f(x,y)dxdy}. \qquad (5)$$

The relations (3) and (5) are usually exploited as necessary conditions for the optimality of A. We will call each A satisfying (3) and (5) a **stationary k-set**. In the next section an iterative procedure for finding stationary k-sets will be described.

3. Lloyd's algorithm. Here we present a widely used method for finding stationary k-sets, published first in Lloyd (1982). In order to find the global optima for W(A) one has to apply the method several times using different initial values for A^0, and then to choose the best among the stationary k-sets obtained. The algorithm consists of the following steps:

1. Start with an arbitrary k-set A^0 .
2. Find a corresponding Voronoi partition $S(A^0)$ using (3).
3. For each $S_i(A^0)$ from $S(A^0)$ calculate its mass centre by (5). Let A^1 be the set of these k centres.
4. Repeat Steps 2 and 3 using A^1 instead of A^0, then A^2, and so on.
5. Stop when

$$\| A^n - A^{n-1} \| < \varepsilon , \qquad (6)$$

with a pre fixed ε.

In our study the following form of (6) was used:

$$\frac{1}{k} \sum^{k} \left(|a_{ix}^n - a_{ix}^{n-1}| + |a_{iy}^n - a_{iy}^{n-1}| \right) < 0.0001 .$$

All the integrals in (5) were calculated numerically by the Simpson's formula.

Theoretical problems related to the convergence of the process described here have been investigated by several authors (see e.g. Kieffer (1982), Sabin and Gray (1986)) In our calculations at most 17 iterations were required to obtain the necessary precision.

4. The results. The results of our study are presented in Table 1. Not only the k-centres but also the probabilities of the corresponding Voronoi regions, the number of iterations and the minimum value for W(A) have been given.

Let us add that our program (being written by the

first author) was tested in the case of k=2 by comparing its results with those obtained by analytic methods (see e.g. Pärna (1986)). We can state complete coincidence of the results.

Table 1

Optimum k-centres for the two-dimensional normal distribution with zero means, σ_x=1.5, σ_y=1, and ρ=0.

Value of k	k-centres (x-,y-coordinates)	Prob($S_i(A)$)	W_k^*
k = 1	a_1= (0.0000, 0.0000)	1.0000	2.5000
k = 2 2 iter-s	a_1= (1.1968, 0.0000) a_2= (-1.1968, 0.0000)	0.5000 0.5000	1.8176
k = 3 5 iter-s	a_1= (0.0000, 0.0000) a_2= (1.8352, 0.0000) a_3= (-1.8352, 0.0000)	0.4590 0.2705 0.2705	1.4279
k = 4 9 iter-s	a_1= (1.9560, 0.0000) a_2= (0.0000, 0.9131) a_3= (-1.9560, 0.0000) a_4= (0.0000,-0.9131)	0.2283 0.2718 0.2283 0.2718	1.0503
k = 5 16 iter-s	a_1= (2.1839, 0.3591) a_2= (-2.1839, 0.3591) a_3= (0.0000, 0.8738) a_4= (-0.8885,-0.8254) a_5= (0.8885,-0.8254)	0.1572 0.1572 0.2666 0.2095 0.2095	0.8903
k = 6 14 iter-s	a_1= (2.4924, 0.0000) a_2= (-2.4924, 0.0000) a_3= (0.8125, 0.8513) a_4= (0.8125,-0.8513) a_5= (-0.8125, 0.8513) a_6= (-0.8125,-0.8513)	0.1144 0.1144 0.1928 0.1928 0.1928 0.1928	0.7608

k = 7 15 iter-s	a_1= (2.5445, 0.0000) a_2= (-2.5445, 0.0000) a_3= (1.0190, 1.0356) a_4= (1.0190,-1.0356) a_5= (-1.0190, 1.0356) a_6= (-1.0190,-1.0356) a_7= (0.0000, 0.0000)	0.1020 0.1020 0.1494 0.1494 0.1494 0.1494 0.1982	0.6671
k = 8 17 iter-s	a_1= (0.7403, 0.0000) a_2= (-0.7403, 0.0000) a_3= (0.0000, 1.3250) a_4= (0.0000,-1.3250) a_5= (2.1319, 0.8738) a_6= (2.1319,-0.8738) a_7= (-2.1319, 0.8738) a_8= (-2.1319,-0.8738)	0.1845 0.1845 0.1273 0.1273 0.0941 0.0941 0.0941 0.0941	0.6020

It is seen in the table that, for all the values of k, the elements of the k-centres are located symmetrically w.r.t. the vertical (y) axis. For the most values of k (all except k=5) the k-centres are symmetric also w.r.t. the x-axis. It is somewhat surprising that the 5-centre does not contain the point (0,0). Actually, it should not be surprising after reminding that the geometrical properties of k-centres depend significantly on the ratio of variances. For example, it can be predicted intuitively that if σ_x exceeds σ_y, say, ten times, then the 5-centre will contain zero point as well. In the case of k=3 this effect is seen already here with the variance ratio of 1.5. If this ratio equals to 1, all the 3-centres will have the shape of equilateral triangles centered at zero, but will not have zero as their element.

References.

Bock, H.-H. (1991). Optimum partitions of R^p. IFCS-91. Proceedings of the Third Conference of the International

Federation of Classification Societies (Edinburgh, August 6-9, 1991), (to be published).

Cuesta, J.A, Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random veriables. Probab. Th. Rel. Fields, **78**, 523-534.

Kieffer, J.C. (1982). Exponential rate of convergence for Lloyd's method I. IEEE Trans. Inf. Th., **28**, 205-210.

Lloyd, S.P. (1982). Least squares quantization in PCM. IEEE Trans. Inf. Th., **28**, 129-136.

Pärna, K. (1986). Calculation of the optimal partition of Gaussian distribution in Hilbert spaces. Trans. Comput. Centre Univ. Tartu, **53**, 38-59 (in Russian).

Pärna, K. (1987). Optimal partition of the probabilistic metric space. Dissertation. Vilnius University Press. Vilnius (in Russian).

Pärna, K. (1990). On the existence and weak convergence of k-centres in Banach spaces. Acta et Commentationes Universitatis Tartuensis, **893**, 17-28.

Sabin, M. J., Gray, R. M. (1986). Global convergence and empirical consistency of the generalized Lloyd algorithm. IEEE Trans. Inf. Th., **32**, 148-155.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 28-34.

ON THE CORRESPONDENCE BETWEEN MATRIX DERIVATIVE AND SOME BASIC NOTIONS OF MULTIVARIATE STATISTICS

Tõnu Kollo
University of Tartu, Tartu, ESTONIA

Keywords: *Matrix derivative, moments and central moments of a random vector, asymptotic distributions*

1. Introduction. In the last ten years "newer matrix algebra" (Rogers (1980), Magnus, Neudecker (1988)) has been systematically used in building up a course on multivariate statistics (Muirhead (1982), Fang (1990), for example). The technique includes Kronecker product of matrices and vec-operator as basic notions. For an harmonious presentation of the theory we also need a matrix derivative. So far the derivative has not been included in the "tools" in monographs and textbooks on multivariate analysis or has been used in the form nonorganically agreeing with different notions of multivariate statistics (Fang, Zhang (1990)). Presenting multivariate analysis via matrices we should like its notions to be direct generalizations of their one-dimensional analogues. It occurs that the key point for getting a natural presentation of the theory is defining moments of a random vector in a proper way. In the paper a variant of the matrix derivative is proposed, which in the author's opinion determines natural relations between such notions as the characteristic function, moments and central moments of a random vector, multivariate Taylor expansion and asymptotic normality.

2. Matrix derivative. In literature there are two most frequently used ways for defining the matrix derivative, the so-called Neudecker's derivative and the MacRae's derivative. If the vec-operator is used in the definition, we call the derivative a Neudecker's one.

DEFINITION 1. *(The Neudecker's derivative) Let elements of $r \times s$-matrix Y be the functions of a $p \times q$-matrix X. $rs \times pq$-matrix $\frac{dY}{dX}$ is the Neudecker's matrix derivative in the region D if the partial derivatives $\frac{\partial y_{kl}}{\partial x_{ij}}$ are continuous in D and*

$$\frac{dY}{dX} = \frac{d}{dvecX} \otimes vec'Y \qquad (1)$$

where

$$\frac{d}{dvecX} = \left(\frac{\partial}{\partial x_{11}} \cdots \frac{\partial}{\partial x_{p1}} \cdots \frac{\partial}{\partial x_{1q}} \cdots \frac{\partial}{\partial x_{pq}}\right)'.$$

It is necessary to point out that the derivative (1) remains the same if we change matrices Y and X into their vector-representations vecY and vecX:

$$\frac{dY}{dX} = \frac{dvecY}{dvecX}.$$

If we use the Kronecker product of matrices to present a matrix derivative, we call it the MacRae's derivative.

DEFINITION 2. *(The MacRae's derivative). Let elements of $r \times s$ matrix Y be the functions of a $p \times q$-matrix X. $rp \times sq$-matrix $\frac{\partial Y}{\partial X}$ is the MacRae's matrix derivative in the region D if the partial derivatives $\frac{\partial y_{kl}}{\partial x_{ij}}$ are continuous in D and*

$$\frac{\partial Y}{\partial X} = Y' \otimes \frac{\partial}{\partial X}. \qquad (2)$$

where

$$\frac{\partial}{\partial X} = \begin{pmatrix} \frac{\partial}{\partial x_{11}} & \cdots & \frac{\partial}{\partial x_{1q}} \\ \cdots & \cdots & \cdots \\ \frac{\partial}{\partial x_{p1}} & \cdots & \frac{\partial}{\partial x_{pq}} \end{pmatrix}.$$

The derivative (1) was introduced in the paper Neudecker (1969) where also the main properties of the derivative can be found (basic properties of $\frac{dY}{dX}$ are also presented in Kollo, Kinkar (1984), for example). In the original paper MacRae

(1974) presented her derivative slightly differently from the definition (2), namely as a Kronecker product

$$Y \otimes \frac{\partial}{\partial X}.$$

In the following it comes out that for our purposes it is reasonable to use equality (2) for defining the derivative. The properties of the derivative $\frac{\partial Y}{\partial X}$ can be obtained straightforwardly from the properties obtained by MacRae (1974) (presented also in Kollo (1991)). In the following table we have the basic properties of the derivatives (1) and (2).

Table 1.

Properties of Neudecker's and MacRae's derivatives.

function $Y=Y(X)$	Neudecker's derivative $\frac{dY}{dX}$	MacRae's derivative $\frac{\partial Y}{\partial X}$
X	I_{pq}	$I_{p,q}$
c $c\underset{\sim}{X}$ const	cI_{pq}	$cI_{p,q}$
X'	$I_{p,q}$	$\mathrm{vec}\, I_p \otimes \mathrm{vec}'\, I_q$
A vec X $A - r\times pq$	A'	
X_d $X - p\times p$	$(I_{p,p})_d$	$(I_{p,p})_d$
$Y + Z$	$\frac{dY}{dX} + \frac{dZ}{dX}$	$\frac{\partial Y}{\partial X} + \frac{\partial Z}{\partial X}$
$Z=Z(Y)$; $Y=Y(X)$	$\frac{dZ}{dX} = \frac{dY}{dX}\frac{dZ}{dY}$	
$Y=AXB$	$B\otimes A'$	$(B'\otimes A')I_{r,q}=I_{s,p}(A'\otimes B')$
$Z=AYB$	$\frac{dZ}{dX} = \frac{dY}{dX}(B\otimes A')$	$(B'\otimes I_p)\frac{\partial Y}{\partial X}(A'\otimes I_q)$
$W=W(Y,Z)$, $Y=Y(X)$; $Z=Z(X)$	$\frac{dW}{dX}=\frac{dW}{dX}\Big\vert_{Z=\mathrm{const}}+\frac{dW}{dX}\Big\vert_{Y=\mathrm{const}}$	$\frac{\partial W}{\partial X}=\frac{\partial W}{\partial X}\Big\vert_{Z=\mathrm{const}}+\frac{\partial W}{\partial X}\Big\vert_{Y=\mathrm{const}}$
$W=YZ$ $Y=Y(X)$ $Z=Z(X)$	$\frac{dW}{dX}=\frac{dW}{dX}(Y\otimes I_m)+\frac{dY}{dX}(I_s\otimes Z')$	$\frac{\partial W}{\partial X}=\frac{\partial W}{\partial X}(Y'\otimes I_p)+\frac{\partial Y}{\partial X}(Z'\otimes I_q)$

The dimensions of the matrices are indicated in the table if there are differences compared with the definitions (1) and (2). In the table $I_{p,q}$ denotes a $pq \times pq$-permutation matrix and A_d is a diagonal matrix having the same main diagonal as A has. On notions of matrix theory see Magnus, Neudecker (1988) or Kollo (1991) if necessary.

As can be seen from the table, the Neudecker's derivative has many advantages compared with the MacRae's one from the point of view of simplicity of calculation. Specially we indicate the first and third lines of the table and point out that the differentiation of a composite function is possible by chain rule in the Neudecker's case only.

2. Moments and central moments. Let X be a random p-vector

$$X = (X_1, \ldots, X_p)'.$$

The k-th moment of X is understood as a set of all the possible k-th order mixed moments between the coordinates of X:

$$E(X_1^{i_1} \cdot \ldots \cdot X_p^{i_p}), \quad \sum_{j=1}^{p} i_j = k.$$

These expectations can be arranged in order in many different ways. To fix one of them as a definition, we are directed by a well-known relation between the characteristic function of random variable X and its moments $m_k(X)$ in the one-dimensional case:

$$\left.\frac{d^k \varphi_X(t)}{dt^k}\right|_{t=0} = i^k m_k(X),$$

where $\varphi_X(t)$ is the characteristic function of random variable X. It is desirable to have the similar equality in the multivariate case, too. The characteristic function of p-vector X has the form

$$\varphi_X(t) = Ee^{it'X}, \qquad t \in \mathbb{R}^p. \tag{3}$$

In multivariate analysis traditionally the first moment is presented as a p-vector and the second central moment as $p \times p$-matrix:

$$EX = \mu; \qquad DX = \Sigma.$$

Let us define the moment of order k of a random p-vector X from the following equality:

$$\left.\frac{\partial^k \varphi_X(t)}{\partial t^k}\right|_{t=0} = i^k \mu_k(X) \tag{4}$$

Expanding the characteristic function (3) into the Taylor series, we easily obtain the following result

THEOREM 1. *If a random p-vector X has the k-th order finite moment, then*

$$\mu_k(X) = E(\underbrace{X \otimes X' \otimes X \otimes \ldots \otimes X'}_{k\text{-factors}}), \tag{5}$$

if k is even and

$$\mu_k(X) = E(\underbrace{X \otimes X' \otimes X \otimes \ldots \otimes X}_{k\text{-factors}}), \tag{6}$$

if k is odd.

The proof of the Theorem is a simple modification of its analog in Kollo (1991), §43. As a corollary from Theorem 1 we get equalities for central moments $\bar{\mu}_k(X)$, taking into account that the k-th moment of $Y = X - \mu$ is the k-th central moment of X:

$$\bar{\mu}_k(X) = E[\underbrace{(X - \mu) \otimes (X - \mu)' \otimes \ldots \otimes (X - \mu)'}_{k\text{-factors}}], \tag{7}$$

if k is even and

$$\bar{\mu}_k(X) = E[\underbrace{(X - \mu) \otimes (X - \mu)' \otimes \ldots \otimes (X - \mu)}_{k\text{-factors}}], \tag{8}$$

if k is odd.

So, if $k = 2m$, the moments $\mu_k(X)$ and $\bar{\mu}_k(X)$ are symmetric $p^m \times p^m$-matrices, if $k = 2m - 1$, the moments are $p^m \times p^{m-1}$-matrices.

Let us remark that if to change MacRae's derivative (2) to Neudecker's derivative (1) in the equality (4) the k-th moment $m_k(X)$ would be a $p \times p^{k-1}$-matrix:

$$m_k(X) = E(X \otimes X' \otimes X' \otimes \ldots \otimes X')$$

(Kollo (1991), §5.2.).

If we compare the expressions of $\mu_k(X)$ from equality (5) and $m_k(X)$, we see that the first and the second moments have the same form but for higher order moments $m_k(X)$ we lose the property of symmetry if k is even.

When we use the MacRae's derivative in the original form

$$Y \otimes \frac{\partial}{\partial X},$$

k-th moment $M_k(X)$ would be a p^k-vector:

$$M_k(X) = E(X \otimes X \otimes \ldots \otimes X).$$

This way of defining notions of multivariate analysis was systematically used by Traat (1986).

4. Asymptotic normality. The classical result on asymptotic normality (see Anderson (1958), for example) states that if for a sequence $\{X_n\}$ the convergence of random p-vectors X_n takes place:

$$\sqrt{n}\,(X_n - a) \xrightarrow{\mathcal{D}} N(0,\Sigma), \quad n \to \infty,$$

where

$$X_n \xrightarrow{P} a,$$

then for $g\colon \mathbb{R}^p \to \mathbb{R}^q$ we have the convergence

$$\sqrt{n}\,[g(X_n) - g(a)] \xrightarrow{\mathcal{D}} N(0,\xi'\Sigma\,\xi), \tag{7}$$

when $n \to \infty$ if $g(x)$ is continuously differentiable in a neighbourhood of $g(a)$ and

$$\xi = \left.\frac{dg(x)}{dx}\right|_{x=a}$$

is a matrix derivative (1).

In most of the cases of using asymptotic normality in statistics we have sample mean X or/and vectorized sample covariance matrix vecS (both depending on the sample size n) in the role of X_n. Asymptotic behaviour of X and vecS is described by the following theorem.

THEOREM 2. *(Parring (1979)). Let* $\mathcal{X} = (X_1,\ldots,X_n)$ *be a sample of size n with the first moments*

$EX_i = \mu$, $DX_i = \Sigma$, $\mu_4(X_i) < \infty$. *Then, if* $n \to \infty$

$$\sqrt{n}\,(X - \mu) \xrightarrow{\mathcal{D}} N(0,\Sigma),$$

$$\sqrt{n}\,(\mathrm{vec}(S - \Sigma) \xrightarrow{\mathcal{D}} N(0,\Pi), \tag{8}$$

where

$$\Pi = \bar{\mu}_4(X_i) - \mathrm{vec}\Sigma\ \mathrm{vec}'\Sigma.$$

If we are interested in the asymptotic distribution of some statistic T(S), say, sample correlation matrix or eigenvalues and eigenvectors of S, we can get it from the convergences (7) and (8). For that we have to find the derivative $\left.\frac{d\mathrm{vec}T(S)}{d\mathrm{vec}S}\right|_{S=\Sigma}$, as follows from the convergence (7).

From the definitions (1) and (2) we have

$$\frac{d\mathrm{vec}T(S)}{d\mathrm{vec}S} = \frac{\partial \mathrm{vec}T(S)}{\partial \mathrm{vec}S} = \frac{dT(S)}{dS} \tag{9}$$

As we see from Table 1, the Neudecker's derivative has one remarkable advantage as compared with the MacRae's derivative - chain rule can be applied for differentiating a composite function in this case. But as MacRae's derivative equals to Neudecker's one by equality (9) in the asymptotic distribution problems, we can also use chain rule in deriving asymptotic distributions.

References

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Fang, K.-T., Zhang, Y.-T. (1990). Generalized Multivariate Analysis. Springer-Verlag, Science Press, Beijing.

Kollo, T. (1991). Matrix Derivative in Multivariate Statistics. Tartu University Press, Tartu (in Russian).

Kollo, T., Kinkar, T. (1984). Matrix derivative with application to block-matrices. Contributions of Computer Centre of Tartu State University, Tartu, 51, 96-107 (in Russian).

MacRae, E.C. (1974). Matrix derivatives with an application to an adaptive linear decision problem. Ann. Statist., 2, 337-346.

Magnus, J.R., Neudecker, H. (1988). Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester.

Muirhead, R.J. (1982). Aspects of Multivariate Statistical Theory. Wiley, New York.

Neudecker, H. (1969). Some theorems on matrix differentation with special reference to Kronecker matrix products. J. Amer. Statist. Assoc. 64, 953-963.

Parring, A.-M. (1979). Calculation of asymptotic characteristics of sample functions. Acta et Commentationes Universitatis Tartuensis, 492, 86-90. (in Russian).

Rogers, S.G. (1980). Matrix Derivatives. Marcel Dekker, New York.

Traat, I. (1986). Matrix calculus for multivariate distributions. Acta et Commentationes Universitatis Tartuensis, 733, 64-85.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 35-45.

BOUNDS IN MULTIVARIATE DEPENDENCE

Samuel Kotz and **Ene-Margit Tiit**

University of Maryland, Maryland, USA
University of Tartu, Tartu, Estonia

Keywords: *Distributions with given marginals, moment of inertia, Frechet bounds*

1. The set-up of the problem.

Let F_1, F_2, ..., F_k be given univariate distribution functions, and let $\Pi = \Pi(F_1, \ldots, F_k)$ be the class of all k-variate distribution functions, having the marginal distribution functions F_1, ..., F_k.

If $k = 2$ then in the class Π exist both the lower and the upper bounds, $H^-(.)$ and $H^+(.)$ correspondingly. These are the so-called Fréchet bounds, see Fréchet (1951), Hoeffding (1940), defined by the following formulae:

$$H^+(x_1,x_2) = \min(F_1(x_1),F_2(x_2)), \tag{1}$$

$$H^-(x_1,x_2) = \max(0, F_1(x_1) + F_2(x_2) - 1). \tag{2}$$

In the case $k>2$ the upper bound $H^+(x_1,\ldots,x_k)$, being an immediate generalization of the Fréchet bound (1), exists,

$$H^+(x_1, \ldots, x_k) = \min_{1\le i\le k}(F_i), \tag{3}$$

5*

see Cuadras (1981), Kemp (1973), Ruiz-Rivas (1979).

But for getting the lower bound in the case k > 2 the generalization of the formula (2) is not possible. It has been shown, see Ruiz-Rivas (1979), Rüschendorf (1985), Tiit (1984, 1986), that in the multivariate case instead of one minimal distribution $2^{k-1} - 1$ different extremal distributions exist.

In this paper we use another approach - we try to construct a uniquely defined distribution in the class $\Pi(F_1,\ldots,F_k)$ that might be defined as the lower bound and could be in some sense the generalization of the distribution $H^-(.)$, defined by the formula (2).

2. Some properties of the lower bound of the set Π.

Here we assume that the given univariate distribution functions (d.f.) F_i are all continuous and equal, $F_i = F$ $(i = 1, \ldots, k)$ and prove some lemmas about the properties of the lower bound of the set $\Pi(F, \ldots, F)$.

We regard here the exchangeable k-variate distributions, defined in the following way.

Let g(.) be a permutation of the components of the k-variate vector, $g: R^k \Longrightarrow R^k$ and $g(1, \ldots, k) = (i_1, \ldots, i_k)$, and let $f(x) = f(x_1, \ldots, x_k)$ be a function of k arguments. If the equation

$$f(x) = f(g(x))$$

holds for every point $x \in R^k$ and for every permutation g(.), then the function f(x) is exchangeable.[1]

Lemma 1. A lower bound for the k-variate distributions from the class $\Pi(F, \ldots, F)$ must be exchangeable.

Proof. Let $H^-(x) = H^-(x_1, \ldots, x_k)$ be a lower bound for $\Pi(F, \ldots, F)$.

If $H^-(x)$ is not exchangeable for every permutation g(.) and every vector $x = (x_1, \ldots, x_k)$, then there exist such x and g, that $H^-(g(x)) = H^-(x)$.

Suppose that

$$H^-(x) < H^-(g(x)). \qquad (4)$$

[1] The concept of exchangeability used here coincides with the concept of permutation symmetry, given in Shaked and Tong (1991) and is wider than the concept of exchangeability, introduced by by Loeve (1963).

Let us denote $g(x) = x' = (x'_1, \ldots, x'_k)$, and define

$$H^*(x) = H^-(g^{-1}(x)),$$

where $g^{-1}(.)$ denotes the inverse permutation of $g(.)$. It is evident that also $H^*(x) \in \Pi(F, \ldots, F)$. Let us regard the d.f. $H^*(x)$ in the point x':

$$H^*(x') = H^-(g^{-1}(x')) = H^-(x).$$

From the inequality (4) follows that

$$H^*(x') = H^-(x) < H^-(g(x)) = H^-(x'),$$

that is,

$$H^*(x') < H^-(x').$$

The last inequality contradicts to the assumption that $H^-(x)$ is the lower bound, and hence the exchangeability of the lower bound is necessary.

Lemma 2. All h-variate ($1 < h < k$) marginal distribution functions of a k-variate exchangeable distribution function $H(x_1, \ldots, x_k)$ are equal.

Proof. Let us take $h = 2$. Let us have the arbitrary indices i,j,g and f, $1 = i,j,g,f = k$, fulfilling the following conditions:

$$i \neq j,\ g \neq f,$$
$$(i,j) \neq (g,f).$$

We must show that the bivariate marginal d.f. $H_{ij}(x,y)$ is equal to $H_{gf}(x,y)$. These bivariate marginal d.f's can be received from the $H(x_1, \ldots, x_k)$ by the general definition:

$$H_{ij}(x, y) = H(\infty, \ldots, \underset{i}{\infty}, x, \infty \ldots, \underset{j}{\infty}, y, \infty, \ldots, \infty).$$

From the exchangeability it follows that the result does not change when the finite arguments will be on the g-th and f-th place instead of the i-th and j-th place, consequently we have:

$$H_{ij}(x, y) = H_{fg}(x, y).$$

For generalizing the result for the case of multivariate marginals the same discussion should be repeated.

From Lemma 2 the following corollary immediately arises.

Corollary 1. All covariances and correlations of the k-variate distribution, having an exchangeable d.f. $H(x_1, \ldots, x_k)$, are equal.

Now we must find the lower bound $\bar{H}(x_1, \ldots, x_k)$ from the set of all exchangeable k-variate d.f's belonging to the set $\Pi(F, \ldots, F)$. One possible approach to the problem is given in the following paragraph.

3. Definition of the lower bound with the help of the moment of inertia.

The upper bound (maximal distribution) having the d.f. (3), is in the case of equal continuous marginal distributions concentrated on the line defined by the main diagonal of the unit cube (see Kotz, Seeger (1991), Tiit (1986)), that is, on the line $(x \mid x_1 = x_2 = \ldots = x_k)$. The lower bound should be concentrated in the subspace being orthogonal to this line. For finding this set of points we use the concept of the moment of inertia about this line and maximize it.

The moment of inertia (MI) is defined by the following formula:

$$MI=\Big(\sum_{i<j}(DX_i + (EX_i)^2 + DX_j + (EX_j)^2 - 2\mathrm{cov}(X_i,X_j) - 2EX_iEX_j)\Big)/k. \qquad (5)$$

If we assume that the marginal distributions are standardized, $EX_i = 0$, $DX_i = 1$, $i=1, \ldots, k$, then we receive:

$$MI =\Big(\sum_{i<j} 2(1 - \mathrm{corr}(X_i,X_j))\Big)/k.$$

and if we also use the assumption about the exchangeability, that means, $\mathrm{corr}(X_i,X_j) = a$, $i,j = 1, \ldots, k$, $i = j$, we get

$$MI =(k(k - 1)\times 2\times(1 - a))/(2k) = (k - 1)(1 - a).$$

To maximize the MI we must minimize the correlation a.

The idea, that the lower bound must be the distribution function, minimizing all correlations, is the natural generalization of the bivariate lower bound which also minimizes the correlation coefficient.

For finding the minimal possible value for the correlation coefficient we use the following lemma.

Lemma 3. Let the correlation matrix $A = (a_{ij})$ have the following form:

$$a_{ij} = \begin{cases} 1, & \text{if } i = j, \\ a, & \text{if } i = j. \end{cases} \tag{6}$$

Then its determinant has the following value:

$$\det A = ((k - 1)a + 1)(1 - a)^{k-1}.$$

Proof follows immediately from the fact that the eigenvalues v_i, $i = 1, \ldots, k$ of the matrix A, defined by formulae (6), are expressed in the following way, see Rao (1965), Tiit (1984):

$$v_1 = (k - 1)a + 1, \quad v_2 = \ldots = v_k = 1 - a.$$

From here also the following corollary arises.

Corollary 1. In the case of k-variate exchangeable distribution the correlation coefficient a must fulfill the following condition:

$$a > -(k - 1)^{-1}. \tag{7}$$

In the extremal case $a = -(k - 1)^{-1}$ we have $\det A = 0$ and hence the k-variate distribution is degenerated into some subspace. As in this case only one eigenvalue equals to zero, this subspace is k - 1-variate. The inequality (7) also follows from the calculations, made for the regression models in Tiit (1983).

4. Lower bound in the class of k-variate normal distributions

For regarding the question of the existence of the lower bound we assume that the given marginal distributions are standardized normal distributions. For our further calculations we also use the expression of the inverse matrix $A^{-1} = (a^{ij})$ of a matrix A, fulfilling the conditions (6), see Rao (1965), Tiit (1983, 1984).

$$a^{ij} = \begin{cases} ((k - 2)a + 1)/((k - 1)a + 1)(a - 1)), & \text{if } i = j, \\ -a/((k - 1)a + 1)(a - 1)), & \text{if } i = j. \end{cases} \tag{8}$$

Let us find the k-variate density function $h(x_1, \ldots, x_k)$ for the

case when the correlation matrix A fulfills the conditions (6).

$$h(x_1, \ldots, x_k) = (2n)^{-k/2}(\det A)^{-0.5} \exp(-0.5\, x'A^{-1}x),$$

where

$$x'A^{-1}x = (((k-1)a+1)\sum x_i^2 - a(\sum x_i)^2)/((k-1)a + 1)(a-1)).$$

Let us regard the convergence of the series of k-variate normal density functions, having the correlation matrix $A_n = (a_{ij}^n)$, fulfilling the conditions (6), where the nondiagonal elements are

$$a_n = -(k-1)^{-1} + n^{-1}. \tag{9}$$

If $n \to \infty$, then $a_n \to -(k-1)^{-1}$ Let us denote $d_n = (k-1)a_n + 1$, then $\lim d_n = 0+$. Using the result, given in Lemma 3, we receive the following expression

$$\lim_{n\to\infty} h(x_1,\ldots,x_k) = C_1(x) d_n^{-0.5} \exp(-C_2 d_n^{-1}(\sum x_i)^2), \tag{10}$$

where $C_1(x)$ and C_2 have a finite limit in $a_n \to -(k-1)^{-1}$.

From the formula (10) we see that

$$\lim_{n\to\infty} h(x_1,\ldots,x_k) = \begin{cases} 0, & \text{if } \sum x_i \neq 0, \\ \infty, & \text{if } \sum x_i = 0. \end{cases} \tag{11}$$

Consequently, the mass of the limiting k-variate distribution is concentrated in the (k-1)-variate subspace $(\bar{x} | \sum x_i = 0)$, see also Kotz, Seeger (1991). This subspace is orthogonal to the line $(\bar{x} | x_1 = \ldots = x_k)$.

Let us denote by $h_a(x_1, \ldots, x_k)$ the projection of the density function $h(x_1, \ldots, x_k)$ (calculated for the case $r_{ij} = a$) in the subspace $(\bar{x} | \sum x_i = 0)$. Let us regard the series of matrices $A_n = (a_{ij}^n)$ and the corresponding normal distributions. From the general properties of the normal distribution we can conclude that

- all projections of the normal distributions (having the correlation matrix A_n) in the subspace $(\bar{x} | \sum x_i = 0)$, have k-1-variate normal distribution, belonging to class $\Pi(F, \ldots, F)$ and being exchangeable,

- the series of these projections converges (in distribution) to a k-1-variate normal distribution, belonging to the class $\Pi(F, \ldots, F)$ and saving the exchangeability.

Let us regard the limiting k-1-variate normal distribution, and find its parameters using the following calculation.

Let us define the following linear transformation for X:

$$X_k = -\sum_{i=1}^{k-1} X_i,$$

$$X^* = T\,X,$$

where the transformation matrix T is defined in the following way:

$$T = \begin{pmatrix} I_{k-1} & 0 \\ 0 & 0 \end{pmatrix},$$

I_n being the identity matrix of order n.

So as X is degenerated, the component $X_k = 0$ identically, and consequently the distribution of X^* is the desired k-1-variate normal distribution. Its covariance matrix A^* can be expressed in the following way

$$A^* = T\,A\,T',$$

where A is the covariance matrix of the initial distribution defined by (6), where $a = (k-1)^{-1}$. Consequently, A^* is the matrix of order k - 1, defined by the equations (6) and having the same values of correlations, as the matrix A.

The same result can be reached also by another way. From the exchangeability it follows that all the correlations of the k-1-variate limiting distribution must be equal to each other, i.e., the (k-1)-order correlation matrix A_o of the limiting distribution must fulfill the conditions (6).

The straightforward calculation of the projection shows that the value of the correlation coefficient a_o of the matrix A_o of order k - 1 is equal to $-(k-1)^{-1}$.

5. Examples.

1. Let us regard the case k = 2 and denote $x_1 = x$, $x_2 = y$. The hyperplane $(\bar{x} \mid \sum x_i = 0)$ is in this case the line $y = -x$. So as $h^-(x,y) = 0$ behind the line (see Fig 1.), we must calculate

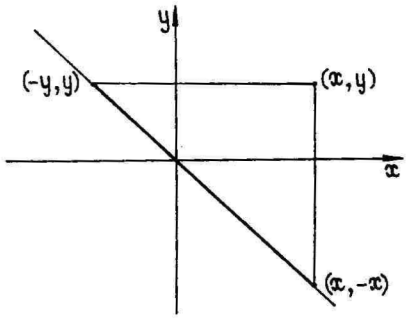

Fig.1.

the value of $H^-(x, y)$ by integrating the function $h^-(x, y)$ on the line segment between the points $(-y, y)$ and $(x, -x)$:

$$H^-(x,y) = \int_{-y}^{x} f(t)dt = F(x) - F(-y) = F(x) + F(y) - 1, \quad \text{if } F(x) > F(-y),$$

$$H^-(x,y) = 0 \quad \text{else.}$$

Hence we received the lower Frechet bound, see (2).

2. Let us regard the case k = 3 and add the notation $x_3 = z$. In this case the upper bound $H^+(x, y, z)$ is the trivariate standardized normal distribution having the correlation matrix A with correlations a = 1. This distribution is concentrated on the line $(\bar{x} \mid x = y = z)$, see Fig. 2 and 3, where the three-dimensional scatterplot of the simulated data (n = 1000 points) of this distribution and its projection on the plane $(\bar{x} \mid z=0)$ are given.

The exchangeable lower bound $H^-(x, y, z)$ is concentrated on the

plane ($\bar{x}$| x + y + z = 0), the minimal possible value of correlations, fulfilling the conditions (6), is a = - 0.5, see Fig. 4, where the three-dimensional scatterplot of the simulated data (n = 500), having the distribution $\bar{H}(x, y, z)$ is represented. The projection of this data set on the plane ($\bar{x}$| z = 0) is given in Fig. 5. As the distribution

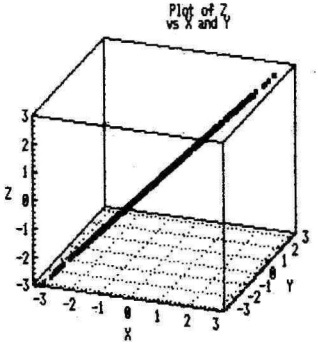

Fig.2.

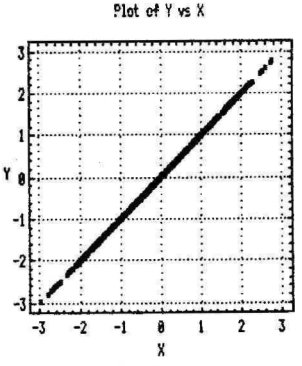

Fig.3.

on this plane is not degenerated, the density function and hence also the distriribution function are positive for all finite points of the plane ($\bar{x}$| z = 0).

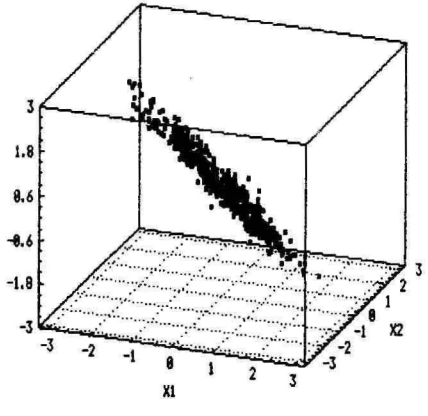

Fig.4.

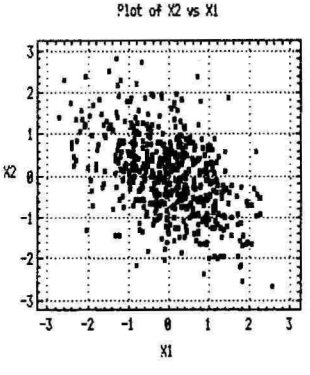

Fig.5.

3. Let us regard also the extremal 3-variate normal distribution,

having the following correlation matrix:

$$R_E = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

Let us denote the d.f. of this distribution by $E(x,y,z)$.

In this case, the density function is concentrated on the line $(\bar{x} \mid x = -y = z)$, see Fig. 6, and its projection on the plane $(\bar{x} \mid z = 0)$ coincides with the line $x = -y$, see Fig. 7. From here we see, that $E(x, y, 0) = 0$ for all points (x, y), satisfying the condition $x + y < 0$. Consequently, for these points the inequality

$$E(x, y, 0) < \bar{H}(x, y, 0)$$

is true.

From here the following corollaries arise:

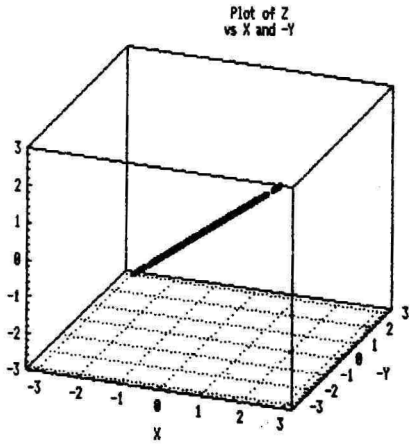

Fig. 6.

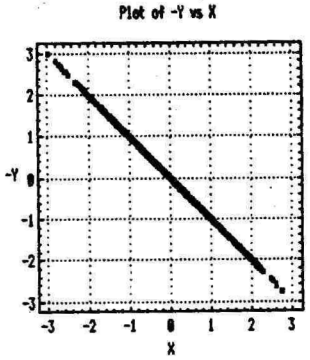

Fig.7.

Corollary 2. For the case $k>2$ the distribution $\bar{H}(x_1, \ldots, x_k)$ is not the lower bound in the proper sense of the word, but it is the lower bound in the set of all exchangeable functions, belonging to the set Π.

From the Corollary 2 and the Lemma 1 also the following result arises:

Corollary 3. For the case k>2 in the set Π the lower bound does not exist.

The authors are thankful to colleagues Anne-Mai Parring and Imbi Traat for the useful suggestions.

References

Cuadras, C.M. (1981) Methods de analisis multivariante. Barcelona, p.52.

Fréchet, M. (1951) Sur les tableaux de correlation dont les marges sont donnees. - Ann. Univ. Lyon. Sect. A, fasc. 14, p. 53 - 77.

Hoeffding, W. (1940) Masstabinvariante Korrelationstheorie. - Schr. Math. Inst. u. Inst. Angew. Math. Univ. Berlin, 5, S. 179 - 233.

Kemp, J.F. (1973) Advanced Problem 5894, Am Math. Monthly, 80, p 83.

Kotz, S., Seeger, J.P. (1991) A new approach to dependence in multivariate distributions. - Distributions with Given Marginals, Dordrecht / Boston / N York, 113 - 127.

Loève, M. (1963) Probability Theory, 3 ed. D. Van Nostrand, Princeton, NJ

Rao, C.R. (1965) Linear statistical inference and its applications, Wiley, New York.

Ruiz - Rivas, C. (1979) Contribuciones a la construccion de la distribuciones multivariantes co marginales fijas. Tesis Doctoral, Fac. Mat, Univ. Barcelona (manuscript).

Shaked, M., Tong, Y.L. (1991) Positive dependence of random variables with a common marginal distribution. - Commun. Statist., Theory, Meth, 20, 4299 - 4313.

Rüschendorf, L. (1985) Construction of multivariate distributions with given marginals. Ann. Inst. Stat. Math, 37, p. 225 - 233.

Tiit, E.A. (1983) Choice of regressors into the linear model. - Zavodskaya Laboratoria, 7, p. 44 - 49 (in Russian).

Tiit, E.-M. (1984) Definition of random vectors with given marginal distributions and given correlation matrix. - Acta et Commentationes Universitatis Tartuensis, 85, p. 21 - 36.

Tiit, E.-M. (1984) Formal Computation of Regression Parameters. - Compstat, Vienna, Physica-Verlag, p. 496 - 502.

Tiit, E.-M. (1986) Random vectors with given arbitrary marginals and given correlation matrix. - Acta et Commentationes Universitatis Tartuensis 733, p 14 - 38.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 46-51.

CONSTRUCTION OF MULTIVARIATE DISTRIBUTIONS WITH GIVEN MULTIVARIATE MARGINALS

J.M.Marco and **C.Ruiz-Rivas**
University of Madrid, Madrid, SPAIN

Keywords: *Copula, fixed marginals, multivariate extremes*

1-INTRODUCTION

The problem of constructing cumulative distribution functions (c.d.f.), F, for a given set of marginals, has a long history starting with Hoeffding(1940) and Fréchet (1951) and can be focused from different points of view:

1.-The study of the class of all possible joint c.d.f. for the given marginals (which is called Fréchet class); here you deal with bounds, and how the degree of dependence is related to the ordering of the distributions.

In the bivariate case, you have the well known Fréchet bounds:

$$F_+(x,y) = \min\{F_1(x), F_2(y)\},$$

$$F_-(x,y) = \max\{F_1(x) + F_2(y) - 1, 0\}, \qquad x, y \in R.$$

that for all bivariate c.d.f., F, with F_1, F_2 as marginals

$$F_-(x,y) \leq F(x,y) \leq F_+(x,y) \text{ for all } x, y \in R$$

The extreme bounds of most of the usual measures of dependence are attained at these distributions (Tchen(1980)) which represent positive and negative functional

This work was partially supported by DGICYT under grant PB88-0178.

relations between the random variables X (distributed as F_1) and Y (distributed as F_2); in fact, if F_1, F_2, are continouos:

a)$F_1(X_1) = F_2(X_2)$ $(a.s)$ iff the joint c.d.f. of (X,Y) is F_+.

b)$F_1(X_1) = 1 - F_2(X_2)$ $(a.s)$ iff the joint c.d.f. of (X,Y) is F_-.

In the multivariate case, given the univariate marginals $F_1,...F_k$ $k > 2$, the upper bound $F_+(x_1,...x_k) = \min\{F_1(x_1),...,F_k(x_k)\}$ $x_i \in R, i = 1,...,k$ is still a c.d.f. ; but the lower bound $F_-(x_1,...x_k) = \max\{F_1(x_1) + ... + F_k(x_k) - (k-1), 0\}$, $x_i \in R$ is not, in general, a proper c.d.f.; it is just a signed measure.

One of the most useful tools for handling multivariate distributions, F,with given univariate marginals, $F_1,...F_k$, is the copula function as named by Schweizer & Sklar (1983) or Uniform representation as named by Kimeldorf & Sampson (1975), which is a cumulative distribution function (c.d.f.) C in the unit cube $[0,1]^k$ with uniform marginals such that given $F_1,...,F_k$ and C:

$$F(x_1,...x_k) = C(F_1(x_1),...,F_k(x_k)) \text{ for all } (x_1,...x_k) \in R^k.$$

Most features of the multivariate structure (dependence structure) are in the copula function which is independent of the marginals and, in general, easier to handle than the original F.

We shall restrict ourselves to continouos F's when C is unique and

$$C(u_1,...,u_k) = F[F^{-1}(u_1),...,F^{-1}(u_k)], \qquad (u_1,...,u_k) \in [0,1]^k$$

Let us denote $C_+(u_1,...,u_k) = \min_i\{u_i\}$ the Fréchet upper bound (which corresponds to a unit mass spread over the main diagonal,$u_1 = u_2 = ... = u_k$, of the unit cube), $C_0(u_1,...,u_k) = \prod_{i=1}^k u_i$ the independence copula and $C_-(u_1,...,u_k) = \max\{u_1 + ... + u_k - k + 1, 0\}$ the Fréchet lower bound which for $k = 2$ is a proper c.d.f. but for $k > 2$ is just a signed measure.

In fact, for $k > 2$, if you spread a unit mass over any of the $2^{k-1} - 1$ remaining diagonals of the unit cube you will get a minimal element of the Fréchet class.

In the multivariate case with given, non overlapping, multivariate marginals (which is the subject of this paper) not even $\min\{F_1(x_1),...,F_k(x_k)\}$, $x_i \in R^{n_i}$, $i = 1,...,k$, is, in general, a proper c.d.f. (Dall'Aglio (1960)).

2.- The study of parametric families of c.d.f. with given marginals (parametric subclasses of the Fréchet class).

There are in the literature, several parametric families of multivariate distributions defined via the corresponding family of copulas (mainly for $k = 2$); let us quote Gumbel(1960), Plackett(1965), Mardia(1970),Johnson&Kotz(1975,1977), Cambanis(1977), Clayton(1978), Frank(1979), Cohen(1980), Genest&Mackay(1986), Marshall&Olkin(1988)....

For a recent treatment of these problems see Dall'Aglio, Kotz & Salinetti(1991).

The purpose of this paper is to grenerate n-variate c.d.f. F with given $F_i, i = 1,...,k, n_i$-variate marginals, $n_1 + ... + n_k = n$. (We do not consider the problem of compatibility of overlaping marginals).

The method we propose is based on an arbitrary k-dimensional copula and needs to impose some restrictions on the marginals. Those restrictions lead, in the simplest case, to max-infinitely divisible marginals .

For the sake of simplicity, we shall consider the case $k = 2$. All the results generalize straightforward for $k > 2$.

We are going to make use of the following

2-PREVIOUS RESULTS.

RESULT 1. *Let F be a n-variate c.d.f. and $\phi : [0,1] \longrightarrow [0,1]$ continouos to the right. Then, $\phi(F)$ is a c.d.f. for all F if :*

$$\begin{aligned} &(i)\ \phi(0) = 0,\ \phi(1) = 1. \\ &(ii)\ \phi^{j)} \geq 0\ 1 \leq j \leq n. \end{aligned} \tag{1}$$

The derivatives here and in the following should be interpreted as distributionals derivatives. The condition $\phi^{j)} \geq 0$ meaning that the j-th distributional derivative is a positive measure.

remarks.

a) if $n = 1$, ϕ could be any univariate c.d.f.

b)if ϕ is absolutely monotone ($\phi^{j)} \geq 0$for all$j$), ϕ can be expanded (Feller(1968)):

$$\phi(t) = \sum_{k=1}^{\infty} p_k t^k$$

and in this case $\phi(F)$ is the c.d.f. of a mixture of maxima; i.e. $\phi(F)$ is the c.d.f. of $(M_1^{(N)}, ..., M_n^{(N)})$ where:

N is a discrete random variable with probability mass fuction $P(N = k) = p_k$, $k = 1, 2, ...$

$M_i^{(N)} = \max_{1 \leq j \leq N} \{X_{ij}\}$, the maxima component wise of a sequence of i.i.d. random vectors $(X_{1j}, ..., X_{nj})$, $j = 1, 2...$ with common joint c.d.f. F.

RESULT 2. *Let G_i, $i = 1, 2$ be n_i – variate c.d.f. ; $\phi : [0,1]^2 \longrightarrow [0,1]$ continouos to the right.*

$\phi(G_1(x_1), G_2(x_2))$, $x_i \in R^{n_i}$ is a c.d.f. in $R^{n_1+n_2}$ for all G_i, $i = 1, 2$ if :

$$\begin{aligned} &i)\phi(1,1) = 1,\ \phi(0,s) = \phi(t,0) = 0 \text{ for all } s,t \in [0,1]. \\ &ii)\frac{\partial^{j+k}\phi}{\partial t^j \partial s^k} \geq 0,\ \ 0 \leq j \leq n_1,\ 0 \leq k \leq n_2. \end{aligned} \tag{2}$$

The marginals being $\phi_1(G_1) = \phi(G_1(x_1), 1)$, $\phi_2(1, G_2(x_2))$.

remarks.

a) If $n_1 = n_2 = 1$, ϕ could be any c.d.f. in the unit square with marginals $\phi_1(t) = \phi(t,1)$, $\phi_2(s) = \phi(1,s)$. Imposing $\phi_i(t) = t$ $i = 1, 2$ ϕ is just any copula.

b) If $n_i \neq 1$ for $i = 1$ or 2, imposing $\phi_i(t) = t$, $i = 1, 2$ you get just the independence case $\phi = \phi_1\phi_2$.

So you cannot generalize the copula with uniform marginals.

3-PROCEDURE

Given F_i, $i = 1,2$ $n_i - variate$ c.d.f., let us suppose that we can express

$$F_i = \phi_i(G_i) \qquad i = 1,2$$

where G_i is a $n_i - variate$ c.d.f. and $\phi_i(t) = t^{\lambda_i}$ $\lambda_i \geq n_i$.

This condition will be satisfied if, for instance, F_i is max-infinitely divisible, (which includes all multivariate extreme value distributions and distributions with Arquimèdean copulas (Genest & Mackay(1986))).

Expanding in Taylor series:

$$\phi_i(t) = \phi_i'(0)t + ... + \frac{\phi_i^{(n_i-1)}(0)}{(n_i-1)!}t^{n_i-1} + \int_0^t \frac{(t-u)^{n_i-1}}{(n_i-1)!}\phi_i^{(n_i)}(u)du =$$

$$= \int_0^t \frac{(t-u)^{n_i-1}}{(n_i-1)!}\phi_i^{(n_i)}(u)du$$

$$t \in [0,1], \qquad i = 1,2$$

As $\phi_i(1) = 1$, we have that

$$\rho_i(u) = \frac{(1-u)^{n_i-1}}{(n_i-1)!}\phi_i^{(n_i)}(u) \qquad u \in [0,1]$$

is a probability density, $i = 1,2$.

Then:

$$\phi_i(t) = \int_o^t \left(\frac{t-u}{1-u}\right)^{n_i-1} \rho_i(u)du \ , \ i = 1,2.$$

Let us note that ρ_i, $i = 1,2$, is the density of a $Beta(\lambda_i - n_i + 1, n_i)$; and if $\lambda_i = n_i$, which is the simplest case, ϕ_i is the c.d.f. of the maximum of n_i i.i.d. Uniform[0,1] random variables, and ρ_i is the density of the minimum.

Now, to built the joint ϕ, let us define:

$$\phi(t,s) = \int_0^t \int_0^s \left(\frac{t-u}{1-u}\right)^{n_1-1} \left(\frac{s-v}{1-v}\right)^{n_2-1} dR(u,v);$$

where $R(u,v)$ is any c.d.f. with ρ_i, $i = 1,2$, as marginals.

It is easy to proof that ϕ satisfies conditions (2), so

$$F(x_1,x_2) = \phi(F_1^{1/\lambda_1}(x_1), F_2^{1/\lambda_2}(x_2)), \qquad x_i \in R^{n_i}, \qquad i = 1,2,$$

is a c.d.f. with F_1, F_2, as marginals.

4-AN EXAMPLE

Let $R(u,v)$ be absolutely continouos with density:

$$\rho(u,v) = n_1 n_2 (1-u)^{n_1-1}(1-v)^{n_2-1}[1+\theta(1-2(1-u)^{n_1})(1-2(1-v)^{n_2})]$$

$$u,v \in [0,1], \qquad \theta \in [-1,1],$$

i.e. the F.G.M. copula.

Then, it is easy to show that

$$\phi(t,s) = \int_0^t \int_0^s \left(\frac{t-u}{1-u}\right)^{n_1-1} \left(\frac{s-v}{1-v}\right)^{n_2-1} dR(u,v) =$$
$$= t^{n_1} s^{n_2}[1+\theta(1-2h_{n_1}(t))(1-2h_{n_2}(s))]$$

where

$$h_n(t) = \int_0^1 n(1-u)^{n-1}(1-tu)^n du = \mathcal{H}(1,-n,n+1,t)$$

is a hypergeometric fuction, which in this case is a Jacobi polynomial

$$h_n(t) = \sum_{k=0}^{n} a_k t^k,$$

with

$$a_k = (-1)^k \frac{(n!)^2}{(n-k)!(n+k)!}.$$

So

$$F(x_1,x_2) = \phi(F_1^{1/n_1}, F_2^{1/n_2}) =$$
$$= F_1(x_1)F_2(x_2)[1+\theta(1-2\sum_{k=0}^{n_1} a_k F_1^{k/n_1}(x_1))(1-2\sum_{j=0}^{n_2} b_j F_2^{j/n_2}(x_2))]$$

$$x_i \in R_i^n, \qquad i=1,2;$$

which can be seen as a generalization of the F.G.M. copula for multivariate marginals.

REFERENCES.

Cambanis, S.(1977).Some properties and generalizations of multivariate Eyraud-Gumbel-Morgenstern distributions. J. Mult.Anal. 7,4,551-59.

Clayton,D.G.(1978).A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65,141-155.

Cohen, L.(1984). Probability distributions with given marginals. J. Math.Phys. 25,2402-2403.

Cuadras C.M. & Auge, J.(1981). A continuous general multivariate distribution and its properties. Comm. Statis. A, 10,339-353.

Dall'Aglio,G.(1960). Les fonctions extremes de la classe de Fréchet à 3 dimensions. Publ.Inst.Stat.Univ. Paris,IX, 175-188.

Dall'Aglio,G..Kotz,S. & Salinetti, . (edit.).(1991). Advances in probability distributions with given marginals. Kluwer.

Feller, W.(1968). An introduction to probability theory and its applications. Vol II. 3rd ed. . Wiley, New York.

Frank, M.J.(1979). On simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$. Aequationes Mathematicae 19, 194-226.

Fréchet, M.(1951). Sur les tableaux de corrélation dont les marges sont données. Annales de l'Université de Lyon, Sciences 4,13-84.

Genest, C. & Mackay, R.J.(1986). Copules arquimédiennes et familles des lois bidimensionelles dont les marges sont données. Canad. J. Statist. 14, 145-159.

Gumbel, E.J.(1960). Distributions des valeurs extremes en plusieurs dimensions. Publ. Inst. Statist. Univ. Paris. 9, 171-173.

Hoeffding, W.(1940). Masstabinvariante Korrelations theorie. Schriftenreihe Math. Inst. Angew. Math. Univ. Berlin 5, 179-233.

Joe, H.(1990). Families of min-stable multivariate exponential and multivariate extreme value distributions. Statist. Probab. Letters 9,75-81.

Johnson, N.L. & Kotz, S.(1975). On some generalized Farlie-Gumbel-Morgenstern distributions. Comm. Statist. 4, 415-427.

Johnson, N.L. & Kotz, S.(1977). On some generalized Farlie-Gumbel-Morgenstern distributions II. Regression, correlation and further generalizations. Comm. Statist. A6, 485-496.

Kimeldorf, G. & Sampson, A.(1975). Uniform representation of bivariate distributions. Comm. Statist. 4, 617-627.

Mardia, K.V.(1970). Families of bivariate distributions. Hafner, Darim.

Marshall, A.W. & Olkin, I.(1988). Families of multivariate distributions. J. Amer. Statist. Assoc. 83, 834-841.

Plackett R.L.(1965). A class of bivariate distributions. J. Amer. Statist. Assoc. 60, 516-522.

Rüschendorf, L.(1990). Fréchet-Bounds and their applications. Angewandte Mathematik und Informatik. Univ. Münster. Tech. rep. .

Schweitzer, R. & Sklar, E.F.(1981). Probabilistic Metric Spaces. Elsevier Science, New York.

Tchen, A.H.(1980). Inequalities for distributions with given marginals. Ann. probab. 8, 814-827.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 52-70.

STATISTICAL ESTIMATION OF GROWING MOLECULAR PHYLOGENETIC TREES

Tõnu Möls

University of Tartu, Tartu, ESTONIA

Summary

Statistical methods for estimating molecular phylogenetic trees are presented, assuming that the topology of the tree is known. A stochastic model of evolving species is presented for generating random phylogenetic binary trees in simulation experiments designed for the investigation of various tree estimates. The results of the experiments suggest that the appending of new present-day species to the tree enables the analyzer to extend the estimation deeper into the past practically without reducing its precision. The results refine our previous paper Möls (to appear).

Keywords: *Molecular evolution, molecular clock, phylogenetic trees, statistical estimation*

1. INTRODUCTION

In the evolutionary theory great progress has been made since the molecular data of evolving species were taken into account, see Curnow and Kirkwood (1989). The principle of the 'Molecular Clock' proposed by E.Zuckerkandl and L.Pauling (1965) makes it possible to construct an objective phylogeny of present-day species from molecular sequences of these species. This principle produces phylogenetic trees not depending on morphometric measurements of the present-day species and fossil founds. On the other hand, general properties of large molecular phylogenetic trees and the corresponding molecular data present certain interest for theoretical biology. For example, a comparison of mutability rates at different periods of evolution offers valuable information about the intimate statistical properties of the process, see Fitch and Langley (1978), Langley and Fitch (1974). Various theoretical concepts of the evolution theory may be verified by comparing the present-day species at the molecular level.

In our earlier paper we have got under some simplifying assumptions the results which suggest that an increase of the number of present-day species allows to extend the estimation of phylogenetic relationships deeper into the past practically without reducing the precision of the estimates. The aim of the present paper is to

refine these results and, in addition, to clarify how effectively molecular sequence data may be used in studying earlier periods of phylogeny if the phylogenetic tree is regular. In a regular tree, branches which correspond to equal time intervals must have the same expected mutational or genetic length.

It appears in a simulation experiment that the situation in regular cases is very similar to those considered in Mõls (to appear). Some minor differences in the problem setting are conditioned by the fact that in the regular case the growth of trees depends on the location of the chronological null (the most ancient moment of time). So we have growing trees with the same topology yet having different chronology which complicates the situation.

Here we will also share with an 'abstract' phylogeny only. It means that the biological identity of the evolving material will be ignored except the most general properties of the evolution process. Our conception includes limitations upon the nature of (abstract) species, the way they mutate and the preliminary knowledge we have. The evolving species will be considered as abstract sequences, and both the topology and the chronology of the phylogenetic tree will be assumed to be known.

The present work was partly supported by the Institute of Zoology and Botany of the Estonian Academy of Sciences and by the Faculty of Mathematics of Tartu University.

2. MAIN ASSUMPTIONS AND NOTATIONS

A phylogenetic tree is a graphic representation of evolutionary relationships between species of a given set of species. If Species 1 has emerged from Species 2 in the course of evolution then Species 2 is more ancient than Species 1 and Species 1 is younger than Species 2. Younger than other species are the modern or present-day species which we call following Curnow and Kirkwood (1989) the OTUs - the Operational Taxonomic Units because they are not necessarily good biological species in a strict sense. The OTUs are graphically represented as the leaves of the tree. All other species correspond to the tree branching points, they are rather theoretical constructions than really existing taxonomic units and we call them the HTUs - the Hypothetical Taxonomic Units.

In this paper we proceed, like in our previous one, on the following assumptions.

(1) The phylogenetic tree is based purely on formal sequences (which may be conceived as some chain-like molecules like enzymes or nucleic acids); both the OTUs and HTUs are represented and compared exclusively by means of these sequences.

(2) Every node or tip of the tree is connected with one or three branches.

(3) Only one type of sequences is present; these sequences have constant length for all species.

(4) The topology or the branching type of the tree is known.

(5) The characteristic sequence is known for every OTU.

(6) All evolutionary changes in the sequence can be thought of as a result of a random series of **attacks**. During any attack, a locus (or position) is chosen in the sequence and the element occupying this locus is replaced by another taken from the set of all possible elements fully randomly (every element is taken with equal probability). Note that an attack may not cause a change, or it may eliminate an earlier change.

(7) All attacks in a time interval occur during relatively short random periods where the mutability of genetic material is very high (we call these activity periods 'mutational explosions').

(8) The number of mutational explosions in a time-interval is random and may be characterized by the Poisson distribution $Po(\alpha)$.

(9) The number of attacks in each mutational explosion is random and may be characterized by the Poisson distribution $Po(\beta)$.

As a result, the total number of attacks in a fixed time-interval is a random variable having the Nyman distribution $Ny(\alpha,\beta)$. Mathematically, the Nymans distribution is characterized by its Moment Generating Function

$$G_{\alpha,\beta}(z) = e^{\alpha\,(e^{\beta(z-1)}-1)} \quad . \qquad (1)$$

The parameter α is referred to as the **local parameter** because it depends on the length of the corresponding evolutionary time-interval (the tree branch length). On the contrary, the parameter β is a **global parameter** because it is assumed to be constant for the entire phylogenetic tree[1].

The aim of the present paper is to estimate the phylogenetic trees and to study the behavior of various estimates if the tree is enlarged by joining new OTUs

---

[1] Negative binomial distribution characterized by Moment Generating Function of the form $G_{\alpha,\beta}(z) = (1 + \beta - \beta z)^{-\alpha}$ is here at least as reasonable as the Neymans distribution. In fact, we have got almost identical results for both distributions.

If one attempts to explain real phylogenetic data by using pure Poisson distribution, he will very soon face contradictions. Sometimes such negative results have given a reason for criticizing the whole molecular clock principle, see Antonov (1986) and Gillespie (1984).

to it. Speaking about the estimation of a tree with given topology, we have the estimation of lengths of its branches in mind. There are three possible ways to characterize the length of a phylogenetic tree branch or, more generally, a path from one node to another. The first measure is the 'expected number of attacks bombing the evolving sequence (or sequences) during its evolution along the branch'. This is an objective and theoretically justified measure which has the advantage of being less disturbed by random fluctuations than other measures. It is of special value for theoretical biology. Unfortunately it has a very abstract meaning.

Secondly, evolutionary time-intervals could be measured by real numbers of attacks which have occurred in them. This way is followed in the next paragraph. The third approach is apparently less abstract and estimates the number of real changes resulting from the underlying series of attacks. Mathematically yet the second way is more preferable than the third because the numbers of attacks behave like additive algebraic values, yet the changes are not additive.

We note that in principle all of these measures are useful when studying complex phylogenetic relations between species. For example, a comparison of real and expected numbers of attacks enables us to evaluate the variance of the number of attacks and just gives a unique statistic for estimating the global parameter ß of the tree.

The concepts described above are symbolized as follows. Let us denote the number of OTUs in the tree by N and the distance between i-th and j-th OTU, $i<j<N$, by D_{ij}. The value of D_{ij} is the minimum number of changes of elements needed to convert the linear structure of the characteristic sequence of the OTU i to that of the OTU j. The column-vector of distances D_{ij}, taken in a lexicographic order, is denoted by D. There are altogether $N(N-1)/2$ elements in D. In fact, distances D_{ij} are random values but for a tree containing specific species they are fixed and known. Further we denote by d_α the distance (number of changes) between the HTUs which correspond to the endpoints of the α-th branch of the tree. Contrary to the D_{ij}, the values of d_α are not known.

We denote by M_{ij} the numbers of attacks which correspond to real changes D_{ij} and by M the corresponding column-vector. The unknown values of M_{ij} can be estimated from the known values of the corresponding D_{ij} as described in the next paragraph. The number of attacks during the evolution period which corresponds to the α-th branch of the tree is denoted by m_α. These values could be estimated.

3. ESTIMATING THE NUMBER OF ATTACKS BETWEEN OTUS

In the present and next paragraph we mainly refine the results of our earlier paper. Let us assume that the evolving sequence has V elements, every element having A possible states (in case of a protein molecule A = 20). After a fixed number x of attacks have happened, minimum 0 and maximum x elements will change their state in the sequence (this is the resulting number of changes remaining in the sequence). An elementary argumentation shows that the expected number Ey of resulting changes is

$$Ey = V(1 - 1/A)(1 - (1 - 1/V)^x) . \qquad (2)$$

If we take $Ey \approx y = D_{ij}$ and $x = M_{ij}$, we can invert this formula for estimating M_{ij}:

$$M_{ij} \approx \begin{cases} \dfrac{\ln\left(1 - \dfrac{D_{ij}}{V(1-1/A)}\right)}{\ln(1-1/V)}, & \text{if } D_{ij} < V(1-1/A) \\ \dfrac{-20}{\ln(1-1/V)}, & \text{if } D_{ij} \geq V(1-1/A) \end{cases} , \qquad (3)$$

where M_{ij} is the number of attacks on the path from OTU i to OTU j to be estimated, and D_{ij} is the observed number of changes. The second term in (3) is rather arbitrary. It is provided for a situation where the random value of D_{ij} replacing the expected number ED_{ij} exceeds the maximum possible value for ED_{ij}[2]. Practically, this limit can be reached only when the number of attacks is very great, a situation which we try to avoid.

Note that the formula (3) can be well improved to correct the rather large bias of the evaluated M_{ij}. Though the correction appears to have only a minor effect on the studied properties of the tree estimates.

4. ESTIMATING THE NUMBERS OF ATTACKS ON ELEMENTARY BRANCHES

Most of the endpoints of the branches in the tree are the unknown HTUs and hence we do not know exactly how many changes or how many attacks correspond to the time intervals. We must estimate these quantities from the information

[2] If a sequence has length V and there are A possibilities for each of its element, no more than V(1-1/A) changes are expected.

contained in vector D_{ij}. Obviously, it is sufficient to investigate only elementary branches which do not include nodes inside it because all other branches are composed of elementary branches.

The situation is explained with the help of Figure 1, where a phylogenetic

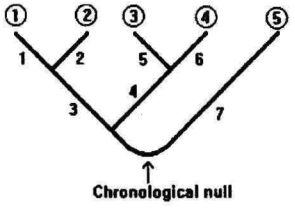

Figure 1. An unrooted phylogenetic tree with 5 present-day species or OTUs (numbers in circles). The branches are numbered for the following use.

tree with 5 OTUs is pictured. Here, for example, the difference D_{13} between OTUs 1 and 3 results from $M_{13} = m_1+m_3+m_4+m_5$ attacks where m_1, m_3, m_4 and m_5 are numbers of attacks on the corresponding elementary branches of the tree. When we are using a special 'design matrix',

$$K = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

we can express the relationship between m_α and M_{ij} in a compact matrix notation as

$$M = Km . \qquad (4)$$

For the estimation of the vector m, various statistical methods can be applied. We have compared some variants of the Maximum Likelihood Method and the Weighted Least Squares Method (for example, taking into account covariances between D_{ij}) and have found in Möls (to appear) that the simple Least Squares

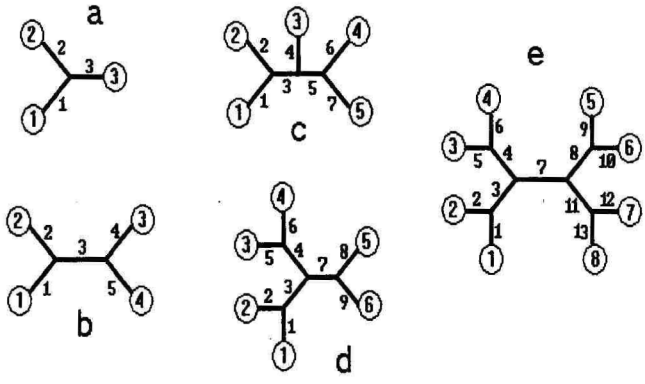

Figure 2. Phylogenetic trees for which the coefficients for estimating the elementary branches are given in Table 1. The branches are numbered as in Table 1.

Method (LSM) is good or even the best:

$$m \approx (K'K)^{-1}K'M \,, \qquad (5)$$

where K' denotes the K transposed and $(K'K)^{-1}$ is the inverse matrix of the regular matrix K'K. Note that numbers d_α of real changes on branches can be estimated by Formula (2) replacing x -> m_α and Ey -> d_α in it.

Examples of linear transformation matrixes $(K'K)^{-1}K'$ for some trees pictured in Figure 2 are given in Table 1. Every row in this matrices, if multiplied by the vector M, gives the estimate of corresponding m_α.

For big numbers N of the OTUs (say, for N > 50), the matrix calculus (5) fails on a PC because of too large dimensions of matrices. For this case, the LSM estimates can be obtained by a special technique developed by us for the experimental investigation of the asymptotic behavior of tree estimates. It works as follows. Suppose we want to estimate m_4, the number of attacks on Branch 4, Figure 1. At first the tree must be redrawn as in Figure 3, making it symmetric by adding new formal sequences 5' and 5" to the mother sequence 5. The Formal Taxonomic Units (FTUs) 5' and 5" are equal to the mother sequence 5 by definition and, consequently, all distances D_{ij} where they figure are equal to the corresponding distances where the mother OTU figures. Note that the needed symmetrization (with respect

Table 1. Coefficients for estimating numbers of attacks on elementary branches of the trees pictured on Figure 2.

Tree	Branch	Matrix $(K'K)^{-1}K$ multiplied by C	C
a	m_1	1 1 -1	2
	m_2	1 -1 1	
	m_3	-1 1 1	
b	m_1	2 1 1 -1 -1 0	4
	m_2	2 -1 -1 1 1 0	
	m_3	-2 1 1 1 1 -2	
	m_4	0 1 -1 1 -1 2	
	m_5	0 -1 1 -1 1 2	
c	m_1	12 4 4 4 -4 -4 -4 0 0 0	24
	m_2	12 -4 -4 -4 4 4 4 0 0 0	
	m_3	-12 6 3 3 6 3 3 -6 -6 0	
	m_4	0 6 -3 -3 6 -3 -3 6 6 0	
	m_5	0 -6 3 3 -6 3 3 6 6 -12	
	m_6	0 0 4 -4 0 4 -4 4 -4 12	
	m_7	0 0 -4 4 0 -4 4 -4 4 12	
d	m_1	4 1 1 1 1 -1 -1 -1 -1 0 0 0 0 0 0 0	8
	m_2	4 -1 -1 -1 -1 1 1 1 1 0 0 0 0 0 0 0	
	m_3	-4 1 1 1 1 1 1 1 1 0 -1 -1 -1 -1 0	
	m_4	0 1 1 -1 -1 1 1 -1 -1 -4 1 1 1 1 0	
	m_5	0 1 -1 0 0 1 -1 0 0 4 1 1 -1 -1 0	
	m_6	0 -1 1 0 0 -1 1 0 0 4 -1 -1 1 1 0	
	m_7	0 -1 -1 1 1 -1 -1 1 1 0 1 1 1 1 -4	
	m_8	0 0 0 1 -1 0 0 1 -1 0 1 -1 1 -1 4	
	m_9	0 0 0 -1 1 0 0 -1 -1 0 -1 1 -1 1 4	
e	m_1	24 4 4 4 4 4 4 -4 -4 -4 -4 -4 -4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	48
	m_2	24 -4 -4 -4 -4 -4 -4 4 4 4 4 4 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
	m_3	-24 6 6 3 3 3 3 6 6 3 3 3 3 0 -3 -3 -3 -3 -3 -3 -3 -3 0 0 0 0 0 0	
	m_4	0 6 6 -3 -3 -3 -3 6 6 -3 -3 -3 -3 -24 3 3 3 3 3 3 3 3 0 0 0 0 0 0	
	m_5	0 4 -4 0 0 0 0 4 -4 0 0 0 0 24 4 4 4 4 -4 -4 -4 -4 0 0 0 0 0 0	
	m_6	0 -4 4 0 0 0 0 -4 4 0 0 0 0 24 -4 -4 -4 -4 4 4 4 4 0 0 0 0 0 0	
	m_7	0 -6 -6 3 3 3 3 -6 -6 3 3 3 3 0 3 3 3 3 3 3 3 3 0 -6 -6 -6 -6 0	
	m_8	0 0 0 3 3 -3 -3 0 0 3 3 -3 -3 0 3 3 -3 -3 3 3 -3 -3 -24 6 6 6 6 0	
	m_9	0 0 0 4 -4 0 0 0 0 4 -4 0 0 0 4 -4 0 0 4 -4 0 0 24 4 4 -4 -4 0	
	m_{10}	0 0 0 -4 4 0 0 0 0 -4 4 0 0 0 -4 4 0 0 -4 4 0 0 24 -4 -4 4 4 0	
	m_{11}	0 0 0 -3 -3 3 3 0 0 -3 -3 3 3 0 -3 -3 3 3 -3 -3 3 3 0 6 6 6 6 -24	
	m_{12}	0 0 0 0 0 4 -4 0 0 0 0 4 -4 0 0 0 4 -4 0 0 4 -4 0 4 -4 4 -4 24	
	m_{13}	0 0 0 0 0 -4 4 0 0 0 0 -4 4 0 0 0 -4 4 0 0 -4 4 0 -4 4 -4 4 24	

to the branch under estimation) is always possible and it is unique if the number of the added FTUs is minimal.

Next, the set of all OTUs and FTUs (but not of HTUs !) is split into the left and right subsets L and R with respect to the Branch 4. Denote by N_L and N_R the numbers of OTUs or FTUs in these sets.

Further, the subset L is a union of upper and lower (with respect to Branch 4) subsets LU and LD. The set R is composed analogously. In the extreme case, L, LU and LD (or R, RU and RD) may consist of one single OTU which is common to all of them. In our example, LU consists of OTUs 1 and 2 and LD consists of FTUs 5' and 5".

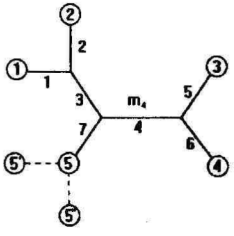

Figure 3. Phylogenetic tree of Figure 1 plotted for the estimation of the number of attacks m_4.

Now the LSM estimate m_4 is obtained from the formula

$$m_4 = \frac{1}{N_L N_R} \sum_{\substack{i \in L \\ j \in R}} M^*_{ij} - \frac{2}{N_L^2} \sum_{\substack{i \in LU \\ j \in LD}} M^*_{ij} - \frac{2}{N_R^2} \sum_{\substack{i \in RU \\ j \in RD}} M^*_{ij} \qquad (6)$$

where

$$M^*_{ij} = \begin{cases} M_{ij}, & \text{if } i<j \\ M_{ji}, & \text{if } i>j \end{cases}. \qquad (7)$$

and sums with empty domain equal to zero. In our example this formula reduces to the following one:

$$m_4 = (M_{13} + M_{14} - 2M_{15} + M_{23} + M_{24} - 2M_{25} - 4M_{34} + 2M_{35} + 2M_{45})/8.$$

The result will be the same as when calculated from the Formula (5) (see Table 1, Tree c, Row m_3, OTUs numbered differently).

Let us briefly discuss the problem of bias. Modeling experiments have shown that the bias of (6) may not be disturbing. For example, the B-series simulation

experiment including $3 \cdot 10^6$ modelled trees with 4, 8, 16, 32, 64 and 128 OTUs (this experiment is explained in the next paragraph) has suggested that the bias of the estimate (6) lies in a 95%-confidence interval of -0.0247 ± 0.040 supposed the expected number of attacks on the corresponding branch is $\alpha\beta = 5$ and the evolving sequence consists of 200 elements with 20 possibilities for each.

5. ASYMPTOTIC RESULTS FOR GROWING TREES

First, let us consider a sequence of special symmetric phylogenetic trees shown in Figure 4. We call this sequence the **A-series**. Every next tree in this sequence has N additional OTUs added to it when compared with the n-th tree having $N = 2^{n+1}$ OTUs. We will assume that all elementary branches have a constant length (in units of evolutionary time) and that the chronological null is located in the central branch. Under these assumptions, the order number $n = \log_2 N - 1$ of the tree represents the depth of the tree and, consequently, the depth increases in the sequence. Note also that we have chosen a sequence of trees with a maximum growth rate.

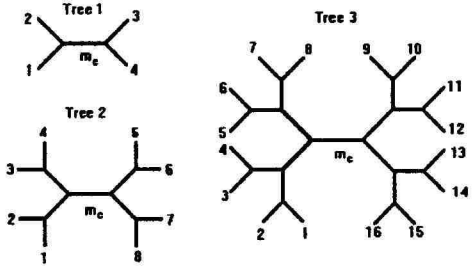

Figure 4. First trees in the A-series of growing symmetric trees. The n-th tree has 2^{n+1} OTUs.

Our main task is to investigate the problem of estimation of the most ancient evolutionary period covered by the central branch of the tree. Intuitively, in this case the accuracy of the estimate depends on two factors having opposite tendencies. First, if the tree grows, the modern species will differ more from its ancestor and

therefore they will contain less information about the ancestor. On the other hand, the number of modern species - the sources of information - grows with the tree exponentially and therefore it is not clear what will happen with the total information when the tree grows - whether it will increase or decrease.

As we are not able to solve this problem analytically, simulation experiments are needed to study the precision of the tree estimates in a line of growing phylogenetic trees. We have studied the central branch estimation with the help of the Formula (6). After taking $N_L = N_R = N/2 = 2^n$, this formula simplifies as follows:

$$m_c = \frac{4}{N^2}\left[\sum_{\substack{i\in L\\ j\in R}} M^*_{ij} - 2\left(\sum_{\substack{i\in LU\\ j\in LD}} M^*_{ij} + \sum_{\substack{i\in RU\\ j\in RD}} M^*_{ij}\right)\right] \qquad (8)$$

where m_c is the estimated total number of attacks bombing the evolving sequence during its evolution period represented by the central branch.

We have run Monte Carlo simulation experiments with A-series phylogenetic trees having N = 4, 8, 16, 32, 64 and 128 OTUs. For each of these six topologies, more than 4000 random sets of OTUs/HTUs (random trees) have been modelled using a chain-like generation mode. According to this mode, one of the HTUs is selected as the starting HTU and its characteristic sequence is initialized to a poly-A 'AAA...AA' with V = 200 occurrences of 'A'. The next HTU (or OTU, if it terminates the branch) is generated by disturbing the previous sequence with a series of attacks, taking the number of attacks randomly according to the Neymans law with local parameter $\alpha = 5$ and global parameter $\beta = 1$.

In this way, on an average $\alpha\beta = 5$ attacks have been generated on each branch of the tree. In every attack, the element in a randomly chosen position of the evolving sequence has been replaced by a random element from the set {A,B,C,...,S,T} of 20 elements. The procedure was continued until the whole tree was generated. Note that the location of the starting OTU and the order of generation have no effect on the final results until we are interested in the differences between the OTUs only.

For every random tree obtained in the way described above, the amount of attacks m_c at the central node has been estimated by the Formula (8) and transformed thereafter with Formula (2) into the estimate of real changes d_c. Comparing the estimates d_c with the corresponding actual numbers of changes recorded immediately in the modelling process, we could empirically estimate the precision (of the type of 'standard deviation') of the estimates d_c.

To get some information about the precision of results of the simulation experiment, all the generated trees have been divided into subsets of equal size

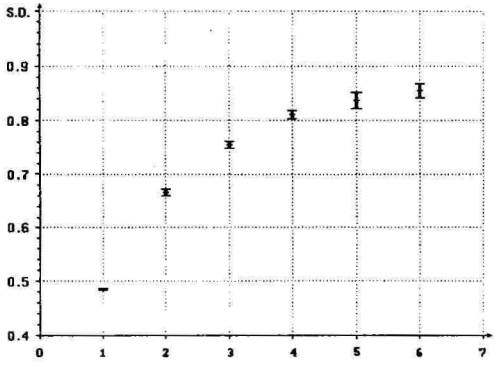

Figure 5. Standard deviations ϵ of the estimated numbers of real changes and their 95% confidence limits for A-series. The X-axis shows the depth of the tree.

K_s each, and in each subset the standard deviation ϵ of the estimate d_c has been found independently:

$$\epsilon = \sqrt{\frac{1}{K_s}\sum_{1}^{K_s}\left(\begin{array}{c}\text{actual number}\\ \text{of changes}\end{array} - \begin{array}{c}\text{estimated number}\\ \text{of changes}\end{array}\right)^2} \quad . \qquad (9)$$

We have used subsets of size K_s = 100 trees, yet the results do not significantly depend on K_s.

The results of the A-series phylogeny modelling experiment are resumed in Figure 5. Here, the X-axis represents the depth of the tree and the Y-axis is for the empirical standard deviation ϵ of the estimate d_c. The depth 1 means 4 modern species or OTUs, the depth 2 means 8 OTUs, etc. Empirical 95% confidence limits of the standard deviation ϵ are shown as well.

As one can see, at first the precision of the estimate decreases simultaneously with the depth of the tree but then it tends to stabilize. This result may be interpreted as follows: a cutback in the precision caused by shifting the evolutionary period into the past can be compensated by the increase of the number of present-day species used in estimation. Of course, depth 6 (128 OTUs) is not sufficient for making clear how this tendency will be continuing when the tree grows further.

In the A-series case, the central branch was the most ancient branch in the tree. Another modelling experiment was undertaken with trees where the chronological null was shifted from the central branch to the right neighbor of it. Let us call

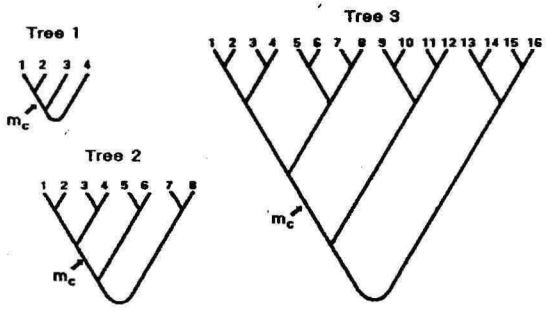

Figure 6. First trees in the B-series of growing symmetric trees with shifted chronological null. The central branch is denoted by m_c.

this sequence the **B-series** of trees, see Figure 6.

In general, the design of the B-series modelling experiment was similar to the A-series experiment[3]. The differences lie mainly in the values of the local parameter which was taken $\alpha = 20$ in B-series for the branch containing chronological null and $\alpha = 10$ for its left neighbor (for all other elementary branches we had $\alpha = 5$). These modifications have been done to get a regular tree.

For each tree type with 4, 8, 16, 32, 64 and 128 OTUs in B-series, 5000 random trees were generated. Results of this experiment are presented in Figure 7. The standard deviations of the estimate m_c are now greater than in Figure 5 because the average distance from the present-day species to the endpoints of the central branch is greater. Nevertheless, the stabilizing tendency can be found in this experiment as well.

---

[3] The generated data were proved by comparing the empirical moments with its theoretical values. The expectation and variance of the number of changes in a branch can easily by found from (1). For example, if two branches have in an average ß' 'explosions' on their separate run and ß" explosions on common part then the covariance of numbers of changes on these branches is

$$\mathrm{cov}(d_1, d_2) = V^2 (1-\tfrac{1}{A})^2 G_1 \left[\tfrac{1}{V}(2G_3 - G_2 - 1) + G_2 - G_3^2\right] + V(1-\tfrac{1}{A}) G_1 (1-G_3)$$

where $G_1 = G_{\alpha,\beta'}(1\text{-}1/V)$, $G_2 = G_{\alpha,\beta''}(1\text{-}2/V)$ and $G_3 = G_{\alpha,\beta''}(1\text{-}1/V)$. In B-series case the theoretical variance for the central branch is 8.821, the corresponding 95%-confidence interval obtained in the modelling experiment was (8.12;10.19).

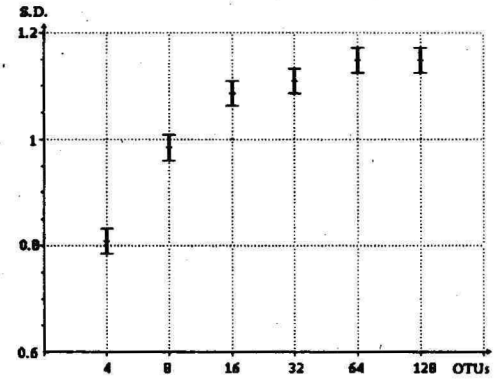

Figure 7. Standard deviations ϵ of the estimated numbers of changes and their 95% confidence limits for B-series simulation experiments.

6. ESTIMATION OF THE EXPECTED NUMBERS OF ATTACKS

When the biological evolution proceeds in a real time and the environment is constant, it leads to a regular phylogeny. The corresponding phylogenetic tree must have fixed chronological null and it must be **regular** in the sense that each two of its branches, if they represent equal time intervals, must have equal expected lengths (expected numbers of attacks). For example, in the tree of Figure 1, the total evolution time of periods 1 and 3 equals to that of periods 4 and 5 and, consequently, on the expectations' level there must hold an equality $m_1+m_3 = m_4+m_5$. The estimate m we have found above does not make use of this fact. Actually m_α estimates a particular realization of a random variable only.

Let us denote by μ a vector analogous to the vector m but satisfying the regularity condition. It may be interpreted as a vector of estimates of the expectations of the lengths of branches and calculated under the assumption that the phylogenetic tree is regular.

Mathematically, the regularity of μ can be expressed by the condition

$$H\mu = 0 \qquad (10)$$

where H is a matrix of constraints. For example, in case of the tree of Figure 1 we have

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix} .$$

Here, for example, the second row means that $\mu_1 + \mu_3 = \mu_4 + \mu_5$.

If the vector m has been estimated earlier, the regular estimate μ can be calculated from the formula

$$\mu = Zm \tag{11}$$

where matrix Z is expressed by known matrices H and K as follows:

$$Z = (1-H^{-}H)[(1-H^{-}H)'K'K(1-H^{-}H)]^{-}(1-H^{-}H)'K'K \ , \tag{12}$$

and m is obtained from (5) or (6). A superscript minus in (12) denotes the generalized inversion of the matrix. Formula (12) enables us to find coefficients for calculating regular estimates from values M_{ij} like those given in Table 1. Some examples of the coefficients of this sort are presented in Table 2.

In general, the Formula (12) leads to rather big matrices. So its use is difficult when the number of OTUs is great. For this situation we suggest a shortcut method analogous to that described in the previous Paragraph. This method produces the estimate (11).

Suppose we want to estimate a branch in a tree with fixed chronological null. Then we have one of two cases.

Case 1: the chronological null is located in the branch α we intend to estimate. Denote by L and R the sets of OTUs connected with the left and right endpoints of this branch, respectively. Further, let N_L and N_R denote the sizes of L and R. Then

$$\mu_\alpha = \frac{1}{N_L N_R} \sum_{\substack{i \in L \\ j \in R}} M^*_{ij} - \frac{1}{2N_{LU}N_{LD}} \sum_{\substack{i \in LU \\ j \in LD}} M^*_{ij} - \frac{1}{2N_{RU}N_{RD}} \sum_{\substack{i \in RU \\ j \in RD}} M^*_{ij} \tag{13}$$

where M^*_{ij} is defined in (7).

Case 2: the chronological null is not located in the branch under estimation. Turn the tree into a position where the estimated branch is horizontal and the chronological null lies right and down from it. For example, the tree in Figure 3 is oriented for estimating the Branch 4, assumed the chronological null is in Branch 6.

Table 2. Coefficients for estimating expected numbers of attacks on elementary branches of the trees pictured on Figure 2. It is assumed that all trees are regular with the chronological null in the marked branch.

Tree	Branch	Matrix $Z(K'K)^{-1}K$ multiplied by C	C
a	m_1	1 0 0	2
	m_2	1 0 0	
	$*m_3$	1 1 1	
b	m_1	6 0 0 0 0 0	12
	m_2	6 0 0 0 0 0	
	m_3	-6 3 0 3 0 0	
	m_4	0 3 0 3 0 0	
	$*m_5$	0 -3 4 -3 4 4	
c	m_1	6 0 0 0 0 0 0 0 0 0	12
	m_2	6 0 0 0 0 0 0 0 0 0	
	m_3	-6 3 0 0 3 0 0 0 0 0	
	m_4	0 3 0 0 3 0 0 0 0 0	
	$*m_5$	0 -3 2 2 -3 2 2 2 2 -6	
	m_6	0 0 0 0 0 0 0 0 0 6	
	m_7	0 0 0 0 0 0 0 0 0 6	
d	m_1	4 0 0 0 0 0 0 0 0 0 0 0 0 0 0	8
	m_2	4 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
	m_3	-4 1 1 0 0 1 1 0 0 0 0 0 0 0 0	
	m_4	0 1 1 0 0 1 1 0 0 -4 0 0 0 0 0	
	m_5	0 0 0 0 0 0 0 0 0 4 0 0 0 0 0	
	m_6	0 0 0 0 0 0 0 0 0 4 0 0 0 0 0	
	$*m_7$	0 -1 -1 1 1 -1 -1 1 1 0 1 1 1 1 -4	
	m_8	0 0 0 0 0 0 0 0 0 0 0 0 0 0 4	
	m_9	0 0 0 0 0 0 0 0 0 0 0 0 0 0 4	
e	m_1	24 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	48
	m_2	24 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	m_3	-24 6 6 0 0 0 0 6 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	m_4	0 6 6 0 0 0 0 6 6 0 0 0 0 -24 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	m_5	0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	m_6	0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	m_7	0 -6 -6 3 3 0 0 -6 -6 3 3 0 0 0 3 3 0 0 3 3 0 0 0 0 0 0 0 0 0 0
	m_8	0 0 0 3 3 0 0 0 0 3 3 0 0 0 3 3 0 0 3 3 0 0 -24 0 0 0 0 0 0
	m_9	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0 0
	m_{10}	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0 0
	$*m_{11}$	0 0 0 -3 -3 4 4 0 0 -3 -3 4 4 0 -3 -3 4 4 -3 -3 4 4 0 4 4 4 4 -24
	m_{12}	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24
	m_{13}	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24

After this adjustment, the following formula can be used:

$$\mu_a = \frac{1}{2N_L N_{RU}} \sum_{\substack{i \in L \\ j \in RU}} M_{ij}^* - \frac{1}{2N_{LD} N_{LU}} \sum_{\substack{i \in LU \\ j \in LD}} M_{ij}^* \qquad (14)$$

where N_{RU} is the number of OTUs in the right upper quadrant on the tree plot.

We have investigated the precision of the Formula (14) in a Monte Carlo experiment with B-series of trees described above. The summary of this experiment is represented by Figure 8. When enlarging the number of OTUs in the tree, the variance of the estimate increases but the tendency seems to fade up.

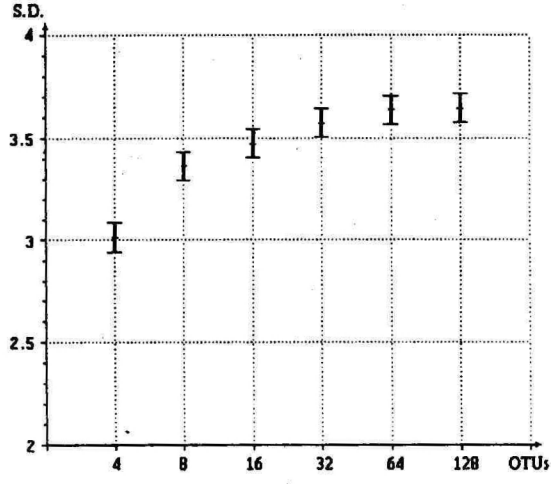

Figure 8. Standard deviations of the estimated expected numbers of attacks with 95% confidence limits for B-series.

The bias of the estimate of (14) has also been studied. The B-series simulation experiment including $3 \cdot 10^6$ modelled trees with 4, 8, 16, 32, 64 and 128 OTUs has suggested that if the Formula (14) is used for estimating the central branch, the bias is in a 95%-confidence interval of -0.0113 ± 0.0418.

Let us mention here another method for the estimation of the expected number of attacks. Note that if we knew the number of attacks in a branch, we could estimate the corresponding expected number of attacks by this value. This idea

leads to an alternative method of estimating the expected numbers of attacks by Formula (5). The precision of this method is illustrated on Figure 9.

The principal difference between the two estimates mentioned is that the second estimate is not a regular one. At the same time, the diagrams of Figure 8 and Figure 9 demonstrate that the second estimate keeps closer to the true value if the number of OTUs is great. This effect may be due to the fact that (5) uses about two times as much OTUs' information as does (14) supposed the tree is big enough (compare, for example, the coefficients in Row 7, Tree e, Tables 1 and 2).

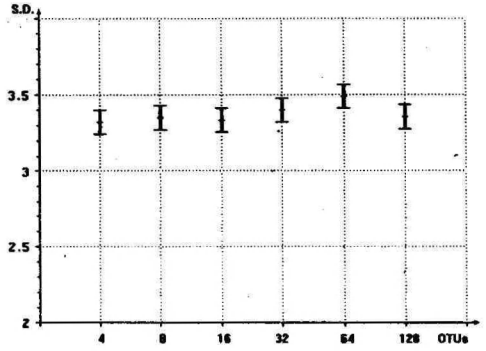

Figure 9. Standard deviations of the second estimate of the expected number of attacks in B-series.

We should like to add some comments to Figure 9. If the exact number of attacks is used for estimating the expected number of attacks on the central branch of a B-series tree, the variance of this estimate would constitute 8.821 (see Footnote 3). Actually we only estimate the exact number of attacks with (5) and this estimate has a variance in the range from 0.73 (for 4 OTUs) to 1.59 (for 64 and 128 OTUs). Assuming that these two errors are independent, we conclude that the variance of our estimate of the expected number of attacks must lie between $0.73+8.82=9.55$ and $1.59+8.82=10.41$. The Figure 9 confirms this argumentation.

7. CONCLUSIONS

The results of our study rise a hypothesis that if the number of the modern species in a phylogenetic tree increases exponentially, the corresponding increase of

the available genetic information will compensate the loss of information of phylogenetic ancestors due to DNA replication errors. This idea concerns both the evaluation of the actual number of molecular changes in early evolutionary periods and the evaluation of expected number of changes. The problem still remains under discussion because only a simulation study has been carried out. Moreover, the answer may drastically depend on the evolution model and its parameters used in theoretical investigations.

We hope that the Formulae (6) and (14) together with the Covariance Formula (see Footnote 3) give a good basis for analytic calculations of the variance of these estimates.

References.

Antonov, A.S. (1986). Evolution of plant genes: a challenge to the molecular clock theory (in Russian). - Priroda, **847**, 68-77.

Curnow, R.N., Kirkwood, T.B.L. (1989). Statistical analysis of deoxyribonucleic acid sequence data - a review. - J.Royal Stat. Soc., Series A, 152(2), 199-220.

Fitch, W.M., Langley, G.H. (1978). Protein evolution and the molecular clock. - Evolution on protein molecules. Tokyo: Japan Scientific Societies, 45-60.

Gillespie, J.H. (1984). Molecular clock may be an episodic clock. - Proc. Nat. Acad.Sci.USA, 81, 8009-8013.

Langley, G.H., Fitch, W.M. (1974). An examination of the rate of molecular evolution. - J.Mol.Evol., 3, 161.

Möls, T. Statistical estimation on molecular phylogenetic trees. - Lectures in theoretical biology (ed. K.Kull, T.Tiivel). London (to appear).

Zuckerkandl, E., Pauling, L. (1965). Evolutionary divergence and convergence in proteins. - Evolving genes and proteins. N.Y.: Acad.Press, 97-166.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 71-75.

BLU ESTIMATION WITH AGGREGATE DATA

Heinz Neudecker
University of Amsterdam, Amsterdam, HOLLAND

Keywords: *Aggregate data of different periodicity, matrix calculus, Lagrange multipliers*

In an interesting contribution Lucke (1990) solved the problem of BLU estimation with aggregate data of different periodicity. For this purpose he employed a powerful projection theorem due to Drygas (1970). The problem can, however, be solved by means of time-honoured techniques, viz. matrix calculus combined with Lagrange multipliers. This will be done in this paper.

Let there be n endogenous unobservable variables and the associated models

$$y_i = X\beta_i + u_i \qquad (i = 1...n)$$

where X is a known (rTxk) nonstochastic matrix of full column rank and the u_i are correlated error vectors with zero means. The vector $u := (u_1' \ ... \ u_n')'$ has positive definite variance $D(u)=\Omega\otimes I_{rT}$. The (nxn) matrix Ω is known.
Available is the following information on the unobservable variables:

1) $y := \Sigma_{i=1...n}\, y_i$

2) $\tilde{y}_i := Sy_i\,, \qquad (i = 1...n)$

where $s_r:=(1... \ 1)'$ is a r-dimensional summation vector and $S:=I_T\otimes s_r'$ is a (TxrT) summation

matrix.

Following Lucke we formulate

THEOREM 1. *Under the assumption that* $|X'S'SX| \neq 0$ *the BLUE of* β_j $(j=1...n)$ *is unique and given by*

$$\hat{\beta}_j := \delta_j/\gamma b_j + (1-\delta_j/\gamma)\tilde{b}_j$$

where

$$\tilde{b}_j := (X'S'SX)^{-1}X'S'\tilde{y}_j$$

$$b_j := (X'X)^{-1}X'y - \Sigma_{i\neq j}\tilde{b}_i$$

$$\gamma := s_n'\Omega s_n \ .$$

$$\delta_j := s_n'\Omega e_j \ .$$

e_j *being the* j^{th} *unit vector of dimension n.*

Its proof follows in a straightforward manner.

PROOF. We consider without loss of generality j=1 and define

$$\Sigma_1 := \Sigma_{i=1...n} \text{ and } \Sigma_2 := \Sigma_{i=2...n} \ .$$

Because $\tilde{y}_1 = Sy - \Sigma_2\tilde{y}_i$ we examine the general linear estimator of β_1

$$d_1 := D_1y + \Sigma_2D_i\tilde{y}_i = D_1\Sigma_1u_i + \Sigma_2D_iSu_i + \Sigma_2D_iSX\beta_i +$$

$$+ D_1X\Sigma_1\beta_i = Gu + \Sigma_2D_iSX\beta_i + D_1X\Sigma_1\beta_i \ ,$$

where $G := s_n'\otimes D_1 + \Sigma_2D_iS(e_i'\otimes I_{rT})$.

Unbiasedness requires $D_1X=I_k$ and $D_iSX=-I_k$ (i=2...n).

The variance of d_1, $D(d_1)$ equals $GD(u)G' = G(\Omega \otimes I_{rT})G'$.
Define then the Lagrangean function

$$\psi := \tfrac{1}{2}\mathrm{tr}G(\Omega \otimes I_{rT})G' - \mathrm{tr}L_1(D_1X - I_k) - \mathrm{tr}\Sigma_2 L_i(D_iSX + I_k) .$$

This leads to the first differential

$$d\psi = \mathrm{tr}(\Omega \otimes I_{rT})G'(s_n' \otimes dD_1) - \mathrm{tr}XL_1dD_1$$

$$+ \Sigma_2 \mathrm{tr}(\Omega \otimes I_{rT})G'(dD_i)S(e_i' \otimes I_{rT}) - \Sigma_2 \mathrm{tr}SXL_i dD_i .$$

The <u>first</u> term can be rewritten as

$$\mathrm{tr}(s_n' \otimes dD_1)(\Omega \otimes I_{rT})G' =$$

$$= \mathrm{tr}(1 \otimes dD_1)(s_n' \otimes I_{rT})(\Omega \otimes I_{rT})(1 \otimes G')$$

$$= \mathrm{tr}(s_n'\Omega \otimes I_{rT})(1 \otimes G'dD_1)$$

$$= \mathrm{tr}(s_n'\Omega \otimes I_{rT})G'dD_1 .$$

The <u>third</u> term can be rewritten as $\Sigma_2 \mathrm{tr}(e_i'\Omega \otimes S)G'dD_i$ in about the same manner.
From this follow the first-order minimum conditions

$$G(\Omega s_n \otimes I_{rT}) = L_1'X' \qquad (1)$$

and

$$G(\Omega e_\ell \otimes S') = L_\ell'X'S' . \qquad (\ell = 2...n) \qquad (2)$$

We replace G by its definition and get

$$\gamma D_1 + \Sigma_2 \delta_i D_i S = L_1'X' \qquad (1^*)$$

$$\delta_\ell D_1 S' + r\Sigma_2 \omega_{i\ell} D_i = L_\ell'X'S' , \qquad (2^*)$$

where $\omega_{i\ell} := e_i'\Omega e_\ell$.

Our next task is to solve for L_1' (by postmultiplying (1^*) by X and using the unbiasedness conditions), insert the result in (1^*), solve for D_1 from the adjusted (1^*), insert that result in (2^*), solve for L_ℓ' (by postmultiplying the adjusted (2^*) by SX and using the unbiasedness conditions) and insert the result in the adjusted (2^*). This yields the equation system

$$\Sigma_2(\omega_{i\ell}-\delta_i\delta_\ell/\gamma)D_i = (\omega_{1\ell}-\delta_1\delta_\ell/\gamma)(X'S'SX)^{-1}X'S' \quad , \qquad (\ell=2...n) \tag{3}$$

as $\quad \Sigma_2(\delta_i\delta_\ell/\gamma-\omega_{i\ell}) = \omega_{1\ell} - \delta_1\delta_\ell/\gamma$.

In matrix notation (3) reads

$$(A\otimes I_k)\,\mathfrak{D} = (a\otimes I_k)(X'S'SX)^{-1}X'S' \quad , \tag{3*}$$

where

$$A := \begin{pmatrix} \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & & \vdots \\ \alpha_{2n} & \dots & \alpha_{nn} \end{pmatrix} \quad ,$$

$$a := (\alpha_{12} \dots \alpha_{1n})' \quad ,$$

$$\alpha_{i\ell} := \omega_{i\ell} - \delta_i\delta_\ell/\gamma \quad ,$$

$$\mathfrak{D} := (D_2' \dots D_n')' \quad .$$

It is obvious that

$$D_i = -(X'S'SX)^{-1}X'S' \qquad (i=2...n) \tag{4}$$

is a solution of (3).

If $|A| \neq 0$ this solution will be unique.

Consider the matrix

$$\mathfrak{N} := \begin{pmatrix} \gamma & \delta' \\ \delta & \Omega_* \end{pmatrix} \quad ,$$

where

$$\Omega_* := \begin{bmatrix} \omega_{22} & \cdots & \omega_{2n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \omega_{2n} & \cdots & \omega_{nn} \end{bmatrix} \quad \text{and} \quad \delta := (\delta_2 \ldots \delta_n)' .$$

The determinant $|\tilde{\Omega}|$ equals $\gamma\,|\Omega_* - (1/\gamma)\delta\delta'| = \gamma\,|A|$, as $A = \Omega_* - (1/\gamma)\delta\delta'$.

As $|\tilde{\Omega}| = |\Omega| \neq 0$ and $\gamma \neq 0$, we establish that $|A| \neq 0$.
From (4) we derive the solution for D_1:

$$D_1 = (\delta_1/\gamma)(X'X)^{-1}X' + (1-\delta_1/\gamma)(X'S'SX)^{-1}X'S'S . \qquad (5)$$

From this follows

$$d_1 = \delta_1/\gamma(X'X)^{-1}X'y + (1-\delta_1/\gamma)(X'S'SX)^{-1}X'S'Sy$$

$$- (X'S'SX)^{-1}X'S'\Sigma_2\tilde{y}_i$$

$$= \delta_1/\gamma(b_1+\Sigma_2\tilde{b}_i) + (1-\delta_1/\gamma)\Sigma_1\tilde{b}_i - \Sigma_2\tilde{b}_1$$

$$= (\delta_1/\gamma)b_1 + (1-\delta_1/\gamma)\tilde{b}_1 . \qquad \square$$

References.

Drygas, H. (1970): *The Coordinate-free Approach to Gauss-Markov Estimation*, Lecture Notes in Operations Research and Mathematical Systems No. 40, Springer Verlag, Berlin.

Lucke, B (1990): *On BLU-estimation with data of different periodicity*. Economics Letters 33, 173-177.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 76-85.

THE ASYMPTOTIC VARIANCE OF REGRESSION COEFFICIENTS

Anne-Mai Parring

University of Tartu, Tartu, ESTONIA

Keywords: *Regression analysis, asymptotic variance matrix, linear regression function, matrix derivative, limiting distribution*

1. The problem. In the present paper we consider the task of statistical regression analysis. As in the most of the real problems we observe the random vector

$$\mathbf{X}_* = (X_0, X_1, \ldots, X_m)'$$

(or denoting $\mathbf{X} = (X_1, X_2, \ldots, X_m)'$, $\mathbf{X}_* = [X_0 : \mathbf{X}']'$) and have as a sample its values

$$\mathbf{x}_{*i} = (x_{i0}, x_{i1}, \ldots, x_{im})',$$

$i = 1, 2, \ldots, n$. We suppose that X_0 is the dependent variable and $X_1, X_2, \ldots, X_m$ are the arguments - the independent variables.

We are looking for the best linear function of arguments

$$l(\alpha_*, \mathbf{X}) = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_m X_m$$

to describe the dependent variable X_0. As usual, for determining that linear function it is suitable to use least-square (LS) condition

$$E(X_0 - l(\alpha_*, \mathbf{X}))^2 = \min_{\alpha_*} E(X_0 - l(a_*, \mathbf{X}))^2. \qquad (1)$$

The linear function $l(\alpha_*, \mathbf{X})$ fulfilling that condition is named linear regression function (LRF).

It is well known that LRF exists if vector $\mathbf{X}_*$ has all the second-order moments. Let us denote $E\mathbf{X}_* = \mu_*$.

$$\mu_* = [\mu_o : \mu']',$$

and $D\mathbf{X}_* = \Sigma_*$.

$$\Sigma_* = \left[\begin{array}{c|c} \sigma_{oo} & \sigma_o' \\ \hline \sigma_o & \Sigma \end{array}\right].$$

From LS condition (1) follows that the coefficients of LRF have to satisfy the system

$$\begin{cases} \alpha_o + \alpha'\mu = \mu_o \\ \Sigma\alpha = \sigma_o \end{cases}$$

and hence (if Σ has full rank),

$$\begin{cases} \alpha_o = \mu_o - \alpha'\mu \\ \alpha = \Sigma^{-1}\sigma_o \end{cases} \qquad (2)$$

The coefficient of determination ρ^2 is given by formula

$$\rho^2 = \frac{\sigma_o' \Sigma \sigma_o}{\sigma_{oo}}.$$

In real problems the mean vector μ_* and variance matrix Σ_* are unknown. We have to use their sample estimates

$$\bar{\mathbf{x}}_* = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{*i}$$

and (3)

$$S_* = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{*i}\mathbf{x}_{*i}' - \bar{\mathbf{x}}_*\bar{\mathbf{x}}_*',$$

which give the estimations of LRF coefficients

$$\begin{aligned} \mathbf{a} &= S^{-1}\mathbf{s}_o \\ a_o &= \bar{x}_o - \mathbf{a}'\bar{\mathbf{x}}. \end{aligned} \qquad (4)$$

As the true LRF coefficients depend only on μ_* and Σ_*, we may choose very different distributions for vector $\mathbf{x}_*$ with the same LRF coefficients.

Example 1. (a) Suppose m=1, so we have

$$\mathbf{X}_* = (X_o, X_1)'.$$

Let the vector $\mathbf{X}_*$ have a discrete distribution with

probability function

$$p(i,j) = P(X_o = i,\ X_1 = j) = \frac{\lambda^j}{i!(j-i)!}\, p^i(1-p)^{j-i}e^{-\lambda},$$

$j = 0, 1, 2, \ldots;\quad i = 0, 1, \ldots, j.$

It is easy to see that the marginal distribution of componenets is Poisson distribution. $X_o \sim P(\lambda p)$ and $X_1 \sim P(\lambda)$. The conditional distribution of X_o by fixed X_1 ($X_1 = j$) is Binomial distribution

$$X_{o.1} \sim B(j,p).$$

Hence the mean vector

$$\mu_* = (\lambda p \quad \lambda)$$

and variance matrix

$$\Sigma_* = \begin{bmatrix} \lambda p & \lambda p \\ \lambda p & \lambda \end{bmatrix}$$

Knowing moments of Poisson and Binomial distributions it is very easy to calculate all the moments of vector X_*. For example

$$m_{o1} = E(X_o X_1) = E(E(X_{o.1})X_1) =$$

$$= \sum_{j=o}^{\infty} j\Big(\sum_{i=o}^{j} iC_j^i p^i(1-p)^{j-i}\Big)e^{-\lambda}\,\frac{\lambda^j}{j!} = p\sum_{j=o}^{\infty} j^2 e^{-\lambda}\,\frac{\lambda^j}{j!} =$$

$$= p(\lambda + \lambda^2)$$

and the central moment

$$\bar{m}_{o1} = m_{o1} - \mu_o\mu_1 = \lambda p.$$

Following the formula (2) we get

$$\alpha_1 = p$$
$$\alpha_o = 0.$$

(b) Let the another vector X_*, $X_* = (X_o, X_1)'$, have the two-dimensional normal distribution with mean vector

$$\mu_* = (\lambda p \quad \lambda)$$

and variance matrix

$$\Sigma_* = \begin{bmatrix} \lambda p & \lambda p \\ \lambda p & \lambda \end{bmatrix}.$$

$X_* \sim N(\mu_*, \Sigma_*)$. Naturally, the LRF coefficients are the same. $\alpha_1 = p$ and $\alpha_o = 0$.

Despite the equality of true values, the distribution of the sample estimates LRF coefficients may be very different. Let at first investigate it by simulating.

In simulating experiment the distributions parameters were chosen in the following way:

$$\lambda = 0.5, \qquad p = 0.6$$

and n - the sample size - was 20. With these parameters 50 samples were generated. For all the samples the LRF coefficients' estimates a_0 and a_1 were calculated. On Fig. 1. there are scattergrams of estimates for discrete and normal distributions. These scattergrams are rather different. May the difference disappair if the sample size increases?

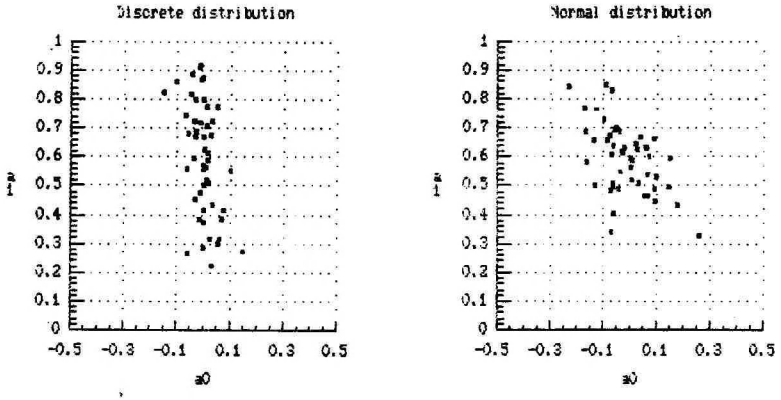

Figure 1.

For the following simulation the sample size n = 100. the distribution parameters are the same. With these parameters 50 samples were generated. On Fig. 2 there are scattergrams of estimates for discrete and normal distribution. They are also very different.

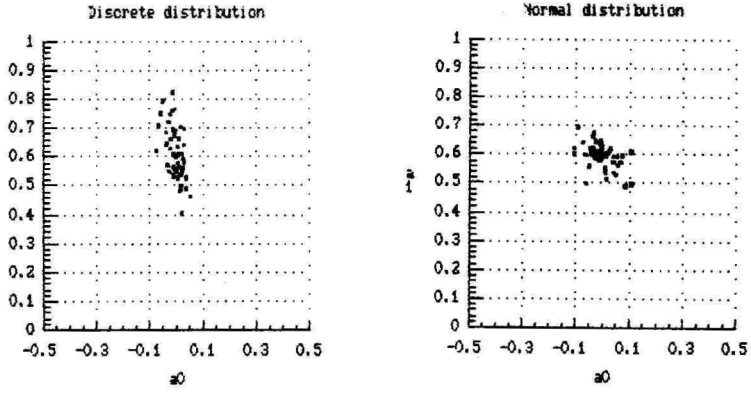

Figure 2.

2. **Asymptotic variance matrix.** Let us try to explain why the behavior of estimates' vectors is so different. It is better to do it using asymptotic distribution.

One quite rough asymptotic approximtion is available by approximate linearization (see, for example, Bandorff-Nilsen, Cox (1989)).

Let $\mathbf{V}_n$, a random vector, converge in probability,

$$\mathbf{V}_n \xrightarrow{P} \mathbf{a},$$

where $\mathbf{a}$ is a constant vector and

$$\sqrt{n}\,(\mathbf{V}_n - \mathbf{a}) \xrightarrow{\mathcal{L}} F,$$

where F is a distribution and $\xrightarrow{\mathcal{L}}$ marks converging in distribution.

Denoting by $\mathbf{V}_*$ a random vector with distribution F, $\mathbf{V}_* \sim F$, we have

$$\sqrt{n}\,(\mathbf{V}_n - \mathbf{a}) \sim \mathbf{V}_* + o(1),$$

where $o(1) \xrightarrow{P} 0$.

Let $\mathbf{h}(\mathbf{V})$ be a twice differentiable vector-function with nonvanishing first derivative at $\mathbf{a}$, $\frac{\partial \mathbf{h}}{\partial \mathbf{v}}(\mathbf{a}) \neq 0$, and with bounded second derivative. Consider the sequence $\{\mathbf{h}(\mathbf{V}_n)\}$ and denote

$$\mathbf{H}_n^* = \sqrt{n}\,(\mathbf{h}(\mathbf{V}_n) - \mathbf{h}(\mathbf{a})).$$

From Taylor's formula follows that

$$H_n^* = \frac{\partial h}{\partial v}(a) V_* + o(1).$$

Denoting the variance matrix of vector V_* with Ξ we get

$$DH_n^* \to \zeta = \frac{\partial h}{\partial v}(a)\, \Xi\, \left(\frac{\partial h}{\partial v}(a)\right)'. \qquad (5)$$

This matrix ζ is named asymptotic variance matrix of $\sqrt{n}\, h(V_n)$.

As we are looking for asymptotic variance matrix of LRF coefficient estimates, it is obvious, that as V_n we have to use vector $[\bar{X}_*' \,\vdots\, \mathrm{vec}_* M_2']'$, where ${}_*M_2$ is matrix of sample second-order moments.

$${}_*M_2 = \frac{1}{n} \sum_{i=1}^{n} X_{*i} X_{*i}'.$$

We dnote ${}_*M_2$ the matrix of second-order moments of vector X_*.

$${}_*M_2 = E(X_* X_*').$$

Hence $\bar{X}_*$ and ${}_*M_2$ give consistent estimates to μ_* and ${}_*M_2$ it is obvious that as a we have to use vector $[\mu_*' \,\vdots\, \mathrm{vec}_* M_2']'$. In such a choise the V_* limiting distribution is normal with asymptotic variance matrix Ξ (Parring (1979a)).

$$\Xi = \left[\begin{array}{c|c} {}_*M_2 - \mu_* \mu_*' & {}_*M_3 - \mu(\mathrm{vec}_* M_2)' \\ \hline {}_*M_3 - \mathrm{vec}_* M_2 \mu' & {}_*M_4 - \mathrm{vec}_* M_2 (\mathrm{vec}_* M_2)' \end{array}\right],$$

where

$${}_*M_3 = E(X_* \otimes X_*' \otimes X_*).$$

$${}_*M_4 = E(X_* \otimes X_*' \otimes X_* \otimes X_*').$$

After quite tedious calculation of the derivative $\frac{\partial \alpha}{\partial v}(\mu_*, \mathrm{vec}_* M_2)$ (see Parring (1979b)) we get the asymptotic variance matrix of LRF estimates

$$\zeta = \left[\begin{array}{c|c} \sigma_{oo}(1 - \rho^2) - 2\mu \cdot \Sigma^{-1} f + \mu \cdot \Pi \mu & f \cdot \Sigma^{-1} - \mu \cdot \Pi \\ \hline \Sigma^{-1} f - \Pi \mu & \Pi \end{array}\right],$$

where

$$\Pi = \Sigma^{-1} C \Sigma^{-1},$$

$$C = (c_{ij}), \quad i, j = 1, 2, \ldots, m,$$

$$c_{ij} = \gamma' ({}_*\bar{M}_4)_{ij} \gamma.$$

$$\gamma = [1 \mid -\alpha']',$$

$$({}_*\bar{M}_4)_{ij} = E((X_i-\mu_i)(X_j-\mu_j)(X_*-\mu_*)(X_*-\mu_*)').$$

$$f = (f_1, f_2, \ldots, f_m)',$$

$$f_i = \gamma'({}_*\bar{M}_3)_i\gamma,$$

$$({}_*\bar{M}_3)_i = E((X_i-\mu_i)(X_*-\mu_*)(X_*-\mu_*)').$$

The asymptotic distribution of regression estimates is normal distribution.

Example 2. (a) It is easy to see that for discrete distribution from Ex.1 the matrices of the third and the fourth central moments are the following

$${}_*\bar{M}_3 = \begin{bmatrix} \lambda p & \lambda p \\ \lambda p & \lambda p \\ \lambda p & \lambda p \\ \lambda p & p \end{bmatrix}$$

and

$${}_*\bar{M}_4 = \begin{bmatrix} \lambda p(1+3\lambda p) & \lambda p(1+3\lambda p) & \lambda p(1+3\lambda p) & \lambda p(1+\lambda(1+2p)) \\ \lambda p(1+3\lambda p) & \lambda p(1+\lambda(1+2p)) & \lambda p(1+\lambda(1+2p)) & \lambda p(1+3\lambda) \\ \lambda p(1+3\lambda p) & \lambda p(1+\lambda(1+2p)) & \lambda p(1+\lambda(1+2p)) & \lambda p(1+3\lambda) \\ \lambda p(1+\lambda(1+2p)) & \lambda p(1+3\lambda) & \lambda p(1+3\lambda) & \lambda(1+3\lambda) \end{bmatrix}$$

Here $\gamma' = (1 \quad -p)'$, $f_1 = \gamma'({}_*\bar{M}_3)_1\gamma = \lambda p(1-p)$ and $c_{11} = \gamma'({}_*\bar{M}_4)_{11}\gamma = \lambda pq(1+\lambda)$. The asymptotic variance matrix for the regression estimates is the following

$$\xi_D = \begin{bmatrix} \lambda & -1 \\ -1 & \dfrac{1+\lambda}{\lambda^2} \end{bmatrix}.$$

(b)The matrices of the third and the fourth central moments for normal distribution from Ex.1 are completely different. As normal distribution is symmetric, ${}_*\bar{M}_3 = 0$. For the matrix of the fourth central moments it is possible to give an expression

$${}_*\bar{M}_4 = \mathrm{vec}\Sigma_*(\mathrm{vec}\Sigma_*)' + \Sigma_*\otimes\Sigma_* + \Sigma_*\circ\Sigma_*,$$

where the operation "$\circ$" is determined in the following way

$$A \cdot B = [(AB)_{ij}],$$
$$(AB)_{ij} = a'_{(i)} \otimes b_{(j)},$$

$a_{(i)}$ - the i-s column of matrix A, $b_{(j)}$ - the j-s column of matrix B. So for multidimensional normal distribution the asymptotic variance matrix of regression estimates (see Parring (1979b)) is

$$\xi = \sigma_{oo}(1-\rho^2)\left[\begin{array}{c|c} 1 + \mu \cdot \Sigma^{-1}\mu & -\mu \cdot \Sigma^{-1} \\ \hline -\Sigma^{-1}\mu & \Sigma^{-1} \end{array}\right].$$

For two-dimensional normal distribution from Ex.1 we get

$$\xi_N = \lambda p(1-p)\begin{bmatrix} 1+\lambda & -1 \\ -1 & \frac{1}{\lambda} \end{bmatrix}.$$

The cause of the different behavior of regression estimates is the presence differences in the third and the fourth central moments of the initial distribution.

3. Analysis of simulation results. To obtain some knowledge how useful the knowing of the asymptotic variance is, two series of simulating experiments were carried out. In the first series 500 samples were generated with 20 elements in each . In the second series 500 samples were generated with 100 elements in each . Denoting the values of LRF coefficients on i-s sample as a_{oi} and a_{1i}, the sample estimates of their standard deviation are calculated as usual

$$\hat{s}_o = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(a_{oi} - \bar{a}_o)^2},$$

$$\hat{s}_1 = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(a_{1i} - \bar{a}_1)^2},$$

(k - number of generated samples).

As it was ealier, $\lambda = 0.5$ and $p = 0.6$. With these values of parameters we get the asymptotic variance matrices

$$\xi_D = \begin{bmatrix} 0.06 & -0.12 \\ -0.12 & 0.72 \end{bmatrix}$$

and

$$\xi_N = \begin{bmatrix} 0.18 & -0.12 \\ -0.12 & 0.24 \end{bmatrix}.$$

Let us denote the standard deviations of a_o and a_1 with s_o and s_1 correspondingly. From definition (5) of asymptotic variance matrix it follows that

$$s_o \simeq \sqrt{\xi_{oo}/n},$$

$$s_1 \simeq \sqrt{\xi_{11}/n}.$$

These values are used as theoretical standard deviations in Table 1. Using these theoretical standard deviations 90%-intervals were calculated . The 90%-interval consists of the sample standard deviation with probability 0.9 if the true distribution of regression coefficient is $N(\alpha_i, s_i)$. As then

$$P(s_i\sqrt{h_{\alpha/2;k-1}/(k-1)} \leq \hat{s}_i \leq s_i\sqrt{\bar{h}_{\alpha/2;k-1}/(k-1)}) = 1-\alpha$$

(k - the number of generated samples, $\hat{s}_i$ - the sample estimate of standard deviation), the limits are given by formulae

$$\underline{l} = s_i\sqrt{h_{\alpha/2;k-1}/(k-1)}$$

$$\bar{l} = s_i\sqrt{\bar{h}_{\alpha/2;k-1}/(k-1)}.$$

($h_{0.05;499} = 439.47$, $\bar{h}_{0.05;499} = 563.37$).

Results of simulating are given in Table 1.

Table 1

Type of density	Sample size	Symbol	Theoret. std.dev.	90% interval		500-smpl est. $\hat{s}_i$
Discrete	20	s_o	0.0548	[0.0514	0.0582]	0.0478
	20	s_1	0.1897	[0.1780	0.2016]	0.1959
Normal	20	s_o	0.0949	[0.0891	0.1008]	0.0980
	20	s_1	0.1095	[0.1028	0.1163]	0.1173
Discrete	100	s_o	0.0245	[0.0230	0.0260]	0.0239
	100	s_1	0.0849	[0.0797	0.0902]	0.0821
Normal	100	s_o	0.0424	[0.0398	0.0451]	0.0429
	100	s_1	0.0490	[0.0460	0.0521]	0.0493

For the same samples were calculated the traditional confidence intervals and the confidence intervals using

sample estimation of asymptotic standard deviation for coefficient α_1 and counted the frequency of capturing the true value. Results are given in Table 2.

Table 2

		Confidence interv. with traditional standard dev.		Confidence interv. with asymptotic standard dev.	
Type of density	Sample size	Freq.	95%-conf. int. for 1-α	Freq.	95%-conf. int. for 1-α
Discrete	20	356	[0.672 0.752]	384	[0.731 0.805]
Normal	20	470	[0.919 0.961]	445	[0.862 0.917]
Discrete	100	372	[0.706 0.782]	454	[0.883 0.933]
Normal	100	474	[0.972 0.957]	472	[0.919 0.96]

As conclusion we have to say:

1) the asymptotic variance matrix is a rough aproximation, it will work if the sample size is large;
2) using the estimation of asymptotic variance in large nonnormal samples may be very useful in sense of correctness of confidence level.

References

Barndorff-Nielsen O.E., Cox D.R. (1989) Asymptotic Techniques for Use in Statistics. Chapman and Hall

Parring A.-M. (1979a) The asymptotic characteristic of some sample functions. Acta et Commentationes Universitatis Tartuensis, 492, 86-90 (in Russian)

Parring A.-M. (1979b) The estimating of linear regression and its asymptotic behavior. Acta et Commentationes Universitatis Tartuensis, 492, 91-99 (in Russian)

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 86-93.

INFLUENTIAL OBSERVATIONS AND A MULTIVARIATE LINEAR MODEL

Dietrich von Rosen

Uppsala University, Uppsala, SWEDEN

Keywords: *Influential observations, neighbourhoods, multivariate linear model*

In this paper some aspects of influential observations are discussed. Relations for derivatives will be presented which are helpful when considering linear models. The idea is to consider neighbourhoods which are discribed with the help of Taylor expansions. As an example the derivation of some results presented by de Gruttola et al. (1987) are simplified and extended.

1. Introduction

Over the years many approaches have been suggested for detecting influential observations. A common approach is the method of deleting an observation. For example when estimating parameters the estimator is recalculated with some observation deleted (e.g. see Cook 1977 and Cook & Weisberg 1981). This is done for each observation in the whole set of observations. Thereafter a "distance measure" (summary diagnostic) is used to order the differences between the estimator and the recalculated estimator. For instance the Euclidean distance. Another approach is to study the linear effect on the estimator after a perturbation of the model (Belsley et al. 1980 p.24, Pregibon 1981). Furthermore, one is not and has not been limited to discuss estimators since there are many other quantities which may be of interest in a statistical problem. Cook (1986), for example, used the likelihood function for a perturbed model.

We are interested in neighbourhoods around a certain point, which in this paper is the value of an estimator. However, it may also be a test, the maximum of the likelihood function, the estimated variance of an estimator or any other statistic. In order to describe the neighbourhood we will make a perturbation and then perform a Taylor expansion. Hence, one may view this approach as a straightforward extension of the one discussed by Pregibon (1981). We will be concentrating on obtaining these neighbourhoods but be very sparse with comments on how to utilize them. In fact, there are many unsolved problems concerning this matter so we just hope that the present paper may be directing for future research, especially for multivariate linear models.

Alterations in one or several observations will be made and we are going to study them when the changes are small. Let $T(X)$ denote some statistic and let $T(X;\epsilon)$ denote the statistic with some perturbation of size ϵ in the underlying model. This report will only present a treatment when ϵ is scalar. Of course matrix values can be alloved but this leads to a slightly more technical treatment. However, in many multivariate situations a matrix valued ϵ will be more realistic. In that case we have to use matrix derivatives (e.g. see Magnus & Neudecker 1988). One example of a situation where ϵ may consist of two components is indicated in Belsly et al. (1980 pp. 35-36). In another context Kollo & Neudecker (1991) used matrix derivatives when expanding eigenvectors.

In order to describe a neighbourhood (around $T_0(X)$) we expand $T(X;\epsilon)$ as a polynom in ϵ;

$$T(X;\epsilon) = T_0(X) + \epsilon T_1(X) + \epsilon^2/2T_2(X) + \cdots$$

where $T_0(X), T_1(X), \ldots$ are some functions in X. This may be viewed as a Taylor expansion around $\epsilon = 0$ but it may also be natural to make expansions around other points, e.g. any $\epsilon \in [0,1]$. However, in principle there are no differences between the different expansions. It is up to the statistician to make a decision about the expansion, depending on the specific circumstances concerning the statistical problems. It is also possible to combine different expansions around different points but this will neither be treated. The main point of the paper is that we claim that each term in the expansion tells us something about the neighbourhood and the influence. In practice we hope, however, that terms of higher order can be neglected but this has to be checked and can not be supposed to hold a priori. The results presented in this report will hopefully help to make a decision about the number of terms which should be used in the expansion. We conclude this section by presenting a model and an estimator which will be considered in section 3.

DEFINITION. *Let $Y_i : p \times 1$, $A_i : p \times q$ $q \leq p$, i=1,2, ...,n, $\beta : q \times 1$ and $\Sigma : p \times p$ p.d.. The $Y_i's$ are independently p-variate normally distributed with an unknown dispersion matrix Σ and $E[Y_i] = A_i\beta$, where the $A_i's$ are known design matrices and β is an unknown parameter vector.* ∎

In the definition $E[\bullet]$ and p.d., respectively, stand for expectation and positive definite. The model can be applied in many situations. For example in repeated measurements analysis and growth curve analysis. The estimator given below is an estimator for β and was treated by de Gruttola et al. (1987). It is an estimator which is obtained when after the first cycle stopping an iteration scheme for obtaining

maximum likelihood estimators i.e., if supposing some full rank conditions,

$$\hat{\beta} = (\sum_{j=1}^{n} A_j' \hat{\Sigma}^{-1} A_j)^{-1} \sum_{j=1}^{n} A_j' \hat{\Sigma}^{-1} Y_j \tag{1.1}$$

$$n\hat{\Sigma} = \sum_{j=1}^{n} (Y_j - A_j\hat{b})(Y_j - A_j\hat{b})' \tag{1.2}$$

$$\hat{b} = (A'A)^{-1}A'Y \qquad A = (A_1', A_2', \ldots, A_n')' \quad Y = (Y_1', Y_2', \ldots, Y_n')'. \tag{1.3}$$

2. Derivatives

The estimator given by (1.1) is a weighted estimator and especially we have to take care of the influence ensued by the weight matrix. In particular we are going to study the following perturbed model where the error vectors e_j are independent with mean zero and dispersion Σ;

$$\begin{aligned} Y_j &= A_j\beta + e_j \qquad j = 1, 2, \ldots, n, \quad j \neq s \\ \sqrt{\epsilon}Y_s &= \sqrt{\epsilon}A_s\beta + e_s \end{aligned} \tag{2.1}$$

although other alternatives are also possible. For example, letting V be a prespecified matrix,

$$Y_s + \epsilon V = A_s\beta + e_s$$

or

$$Y_s = G_s\beta + e_s \quad \text{where} \quad G_s = A_s + \epsilon V.$$

To study influence of the estimator we see from (1.1) that we need expansions for a product of matrices and for the inverse of a non-singular matrix. Furthermore, we present some results for the determinant of a non-singular matrix, for eigenvalues and for eigenvectors but these will not be utilized. However, if discussing test statistics the results would appear to be useful.

Let $\partial^k[\bullet]$ denote the *kth* derivative of a matrix with respect to the scalar ϵ where the derivative is defined elementwise and $\partial^0[A] = A$. In this section the derivatives are not evaluated at a specific point whereas in the next section the derivatives will be evaluated at $\epsilon = 1$. The notion of a g-inverse will also be used, i.e. the g-inverse G^- is any matrix satisfying $GG^-G = G$. Furthermore, let $|\bullet|$ stand for the determinant and $y \sim N_n(\bullet, \bullet)$ means that y is normally distributed.

Lemma 2.1. *Suppose that all included matrices are of proper sizes.*

(i) $$\partial^k[AB] = \sum_{i=0}^{k} \binom{k}{i} \partial^i[A]\partial^{k-i}[B]$$

(ii) *Let A be non-singular.*

$$\partial^k[A^{-1}] = -A^{-1}\partial^k[A]A^{-1} + \sum_{j=1}^{k-1}(-1)^{j+1} \sum_{i_1=j}^{k-1} \sum_{i_2=j-1}^{i_1-1} \cdots \sum_{i_j=1}^{i_{j-1}-1} \binom{k}{i_1}\binom{i_1}{i_2}\binom{i_2}{i_3}\cdots\binom{i_{j-1}}{i_j} \times$$

$$A^{-1}\partial^{i_j}[A]A^{-1}\partial^{i_{j-1}-i_j}[A]A^{-1} \times \cdots \times A^{-1}\partial^{i_1-i_2}[A]A^{-1}\partial^{k-i_1}[A]A^{-1} \qquad i_0 = k \geq 1$$

Furthermore, let A be p.d.. If $\partial^r[A] > 0$ for odd r and $\partial^r[A] < 0$ for even r then $\partial^r[A^{-1}] < 0$ for odd r and $\partial^r[A^{-1}] > 0$ for even r. Vice versa is also true. The inequalities are interpreted according to p.d. (Löwner ordering).

(iii) *Let $A : n \times n$ be non-singular and $y \sim N_n(0, (AA')^{-1})$.*

$$\partial^k[|A|^{-1}] = |A^{-1}| \sum_{i_1=0}^{k-1} \sum_{i_2=0}^{i_1-1} \cdots \sum_{i_k=0}^{i_{k-1}-1} \binom{k-1}{i_1}\binom{i_1-1}{i_2}\binom{i_2-1}{i_3}\cdots\binom{i_{k-1}-1}{i_k} \times$$
$$E[y'\partial^{k-i_1}[-1/2AA']yy'\partial^{i_1-i_2}[-1/2AA']y \times \cdots \times y'\partial^{i_{k-1}-i_k}[-1/2AA']y] \quad k \geq 1$$

(iv) *Let λ be a unique eigenvalue to H and Γ the corresponding normalized eigenvector where $\Gamma'\Gamma = 1$ and let $V = H - \lambda I$.*

$$\partial^k[\lambda] = \Gamma' \sum_{i=1}^{k-1} \binom{k}{i} \partial^i[V]\partial^{k-i}[\Gamma] + \Gamma'\partial^k[H]\Gamma$$

$$\partial^k[\Gamma] = -V^-\partial^k[V]\Gamma + \sum_{j=1}^{k-1}(-1)^{j+1} \sum_{i_1=j}^{k-1} \sum_{i_2=j-1}^{i_1-1} \cdots \sum_{i_j=1}^{i_{j-1}-1} \binom{k}{i_1}\binom{i_1}{i_2}\binom{i_2}{i_3}\cdots\binom{i_{j-1}}{i_j} \times$$
$$V^-\partial^{k-i_1}[V]V^-\partial^{i_1-i_2}[V]V^- \times \cdots \times V^-\partial^{i_{j-1}-i_j}[V]V^-\partial^{i_j}[V]\Gamma \qquad k \geq 1$$

PROOF: The statement in (i) can immediately be verified using induction arguments. In (ii) we will also use induction. It is clear that the statement is true for $k = 1$ and $k = 2$. Now, apply (i) to $\partial^k[A^{-1}A]$ and suppose that the statement is valid up to $k - 1$. Then

$$\partial^k[A^{-1}] = -A^{-1}\partial^k[A]A^{-1} + \sum_{r=1}^{k-1} \binom{k}{r} A^{-1}\partial^r[A]A^{-1}\partial^{k-r}[A]A^{-1} +$$
$$\sum_{r=2}^{k-1}\sum_{j=1}^{r-1}\sum_{i_1=j}^{r-1} \cdots \sum_{i_j=1}^{i_{j-1}-1} (-1)^j \binom{k}{r}\binom{r}{i_1}\cdots\binom{i_{j-1}}{i_j} A^{-1}\partial^{i_j}[A]A^{-1} \times \cdots \times A^{-1}\partial^{r-i_1}[A]A^{-1}.$$

Interchanging summation, i.e. using $\sum_{j=1}^{k-2}\sum_{r=j+1}^{k-1}$ instead of $\sum_{r=2}^{k-1}\sum_{j=1}^{r-1}$, and then doing some calculations verifies the relation in (ii). Furthermore, if B and C are p.d. then $BC + (BC)'$ is also p.d.. Thus, by studying

$$A^{-1}\partial[A]A^{-1}\partial[A]\cdots\partial[A]A^{-1}$$
$$A^{-1}\partial^2[A]A^{-1}\partial[A]\cdots\partial[A]A^{-1}$$
$$A^{-1}\partial^2[A]A^{-1}\partial^2[A]\cdots\partial^2[A]A^{-1}$$
etc.

we may show the correspondence between $\partial^r[A] > 0$ and $\partial^r[A^{-1}] < 0$ given in (ii).

The statement in (iii) follows by noting that

$$\begin{aligned}\partial^k[|A|^{-1}] &= \partial^k[(2\pi)^{-n/2}\int_{R^n}\exp(-1/2y'AA'y)dy] \\ &= (2\pi)^{-n/2}\int_{R^n}\sum_{i_1=0}^{k-1}\binom{k-1}{i_1}\partial^{k-i_1}[-1/2y'AA'y]\partial^{i_1}[\exp(-1/2y'AA'y)]dy \\ &= (2\pi)^{-n/2}\int_{R^n}\sum_{i_1=0}^{k-1}\sum_{i_2=0}^{i_1-1}\binom{k-1}{i_1}\binom{i_1-1}{i_2}\partial^{k-i_1}[-1/2y'AA'y]\partial^{i_1-i_2}[-1/2y'AA'y]\times \\ &\quad \partial^{i_2}[\exp(-1/2y'AA'y)]dy \quad \cdots\cdots \ .\end{aligned}$$

For the proof of $\partial^k[\lambda]$ in (iv) we use that $H\Gamma = \lambda\Gamma$. By applying (i) to this relation we obtain

$$\begin{aligned}&H\partial^k[\Gamma] + \partial^k[H]\Gamma + \sum_{i=1}^{k-1}\binom{k}{i}\partial^i[H]\partial^{k-i}[\Gamma] \\ &= \sum_{i=1}^{k-1}\binom{k}{i}\partial^i[\lambda]\partial^{k-i}[\Gamma] + \lambda\partial^k[\Gamma] + \partial^k[\lambda]\Gamma\end{aligned}$$

and if multiplying with Γ' the statement follows. The proof of $\partial^k[\Gamma]$ is a copy of the proof of (ii) when utilizing that $V\Gamma = 0$. ∎

REMARKS: The expectation in (iii) can be obtained in several ways. For example using the results in von Rosen (1988)

$$\begin{aligned}&E[y'\partial^{k-i_1}[AA']yy'\partial^{i_1-i_2}[AA']y \times \cdots \times y'\partial^{i_{k-1}-i_k}[AA']y] \\ &\quad = \sum_{j_0=2}^{2k}\sum_{j_1=2}^{2k-2}\sum_{j_2=2}^{2k-4}\cdots\sum_{j_{k-1}=2}^{2}\overset{k}{\otimes}vec((AA')^{-1})'H(n,k,j_0,\ldots,j_{k-1})'\times \\ &\qquad vec(\partial^{k-i_1}[AA'])\otimes vec(\partial^{i_1-i_2}[AA'])\otimes\cdots\otimes vec(\partial^{i_{k-1}-i_k}[AA']) \quad (3.1)\end{aligned}$$

where $\otimes$ stands for the right Kronecker product, $\overset{k}{\otimes}A = \underbrace{A\otimes\cdots\otimes A}_{k\,times}$, vec is the vec-operator,

$$H(n,r,i_0,\ldots,i_{r-1}) = \prod_{k=0}^{r-1}(I_{n^{2k}}\otimes P(n,i_k,2r-2k))$$

and the permutation matrix

$$P(a,b,c) = I_a\otimes K_{a^{b-2},a^2}\otimes I_{a^{c-b}}$$

where $K_{a,b}$ is the commutation matrix as in Magnus & Neudecker (1979). Moreover, by combining (i), (ii) and (iii) we can write down $\partial^k[|A|^r]$ for arbitrary r. When S is p.d. we may let $S = AA'$ and then it is easy to obtain results for $\partial^k[|S|^r]$ from the above formula.

3. Application

In this section we are going to show how the results of the previous section may be utilized. We have chosen to discuss the model given by (1.1) when a perturbation of the form (2.1) has been made. The results extend as well as simplifies the derivation of some results which have been obtained by de Gruttola et al. (1987). All derivatives will from now on be evaluated at $\epsilon = 1$. Although we are interested in the Taylor expansion we just present the derivatives and leave it to the reader to form a Taylor series.

LEMMA 3.1. *Let $\hat{b}$ be given by (1.3). If (2.1) holds;*

$$\partial^k[\hat{b}] = k!(-1)^k D^{k-1}(A'A)^{-1}A_*'e_0$$

where $D = (A'A)^{-1}A_*'A_*$, $D^k = \underbrace{DD \times \cdots D}_{k \text{ times}}$ and the residual $e_0 = Y_* - A_*(A'A)^{-1}A'Y$.

PROOF: Since $\partial^r[A'Y] = 0$ if $r > 1$ we obtain

$$\partial^k[\hat{b}] = \partial^k[(A'A)^{-1}]A'Y + k\partial^{k-1}[(A'A)^{-1}]A_*'Y_*$$

and since

$$\partial^k[(A'A)^{-1}] = k!(-1)^k D^k(A'A)^{-1}$$

the lemma is verified with the help of some calculations. ∎

REMARK: It is worth observing that the lemma gives an arbitrary expansion of a least squares estimator and that in many situations the series will be a decreasing alternating series and hence error bounds are easily obtained.

LEMMA 3.2. *Let $\hat{\Sigma}$ be given by (1.2). If (2.1) holds;*

$$\partial[n\hat{\Sigma}] = (Y_* - A_*\hat{b})(Y_* - A_*\hat{b})' + \sum_{j=1}^{n}(-(Y_j - A_j\hat{b})\partial[\hat{b}']A_j' - A_j\partial[\hat{b}](Y_j - A_j\hat{b})'),$$

$$\partial^k[n\hat{\Sigma}] = -kA_*\partial^{k-1}[\hat{b}]Y_*' - kY_*\partial^{k-1}[\hat{b}']A_*' + kA_*\sum_{i=0}^{k-1}\binom{k-1}{i}\partial^i[\hat{b}]\partial^{k-1-i}[\hat{b}']A_*' +$$

$$\sum_{j=1}^{n}(-A_j\partial^k[\hat{b}]Y_j' - Y_j\partial^k[\hat{b}']A_j' + A_j\sum_{i=0}^{k}\binom{k}{i}\partial^i[\hat{b}]\partial^{k-i}[\hat{b}']A_j'). \qquad k > 1$$

PROOF: The proof follows by applying straightforward calculations to (1.2) and observing that the *sth* observation has to be treated separately. ∎

Finally we present a result for $\hat{\beta}$. However the situation is still more complex since we need derivatives for $\hat{\Sigma}^{-1}$ and therefore no details are given and for practical use we have to rely on a computer. The proof will be omitted since it is similar to the one in lemma 3.2.

THEOREM 3.1. *Let $\hat{\beta}$ be given by (1.1). If (2.1) holds;*

$$\partial^k[\hat{\beta}] = \sum_{i=0}^{k} \binom{k}{i} \partial^i[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]\partial^{k-i}[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j]$$

with

$$\partial^r[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j] = rA_*'\partial^{r-1}[\hat{\Sigma}^{-1}]A_* + \sum_{j=1}^{n} A_j'\partial^r[\hat{\Sigma}^{-1}]A_j$$

$$\partial^r[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j] = rA_*'\partial^{r-1}[\hat{\Sigma}^{-1}]Y_* + \sum_{j=1}^{n} A_j'\partial^r[\hat{\Sigma}^{-1}]Y_j$$

and $\partial^i[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]$ is obtainable from lemma 2.1 (ii). ∎

COROLLARY 3.1.

$$\partial[\hat{\beta}] = (\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}\partial[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j] + \partial[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j$$

$$\partial^2[\hat{\beta}] = (\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}\partial^2[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j] + 2\partial[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]\partial[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j] +$$
$$\partial^2[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}Y_j$$

with

$$\partial[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]$$
$$= (\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}\partial[\hat{\Sigma}]\hat{\Sigma}^{-1}A_j - A_*'\hat{\Sigma}^{-1}A_*)(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1},$$

$$\partial^2[(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}]$$
$$= -(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}\partial^2[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j](\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1} +$$
$$2(\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}\partial[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j](\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}\partial[\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j](\sum_{j=1}^{n} A_j'\hat{\Sigma}^{-1}A_j)^{-1}$$

and

$$\partial[n\hat{\Sigma}] = e_0e_0' + \sum_{j=1}^{n}(Y_j - A_j\hat{b})e_0'A_*(A'A)^{-1}A_j' + A_j(A'A)^{-1}A_*'e_0(Y_j - A_j\hat{b})',$$

$$\partial^2[n\hat{\Sigma}] = 2e_0e_0'A_s(A'A)^{-1}A_s' + 2A_s(A'A)^{-1}A_s'e_0e_0' -$$
$$2\sum_{j=1}^{n}(Y_j - A_j\hat{b})e_0'A_s(A'A)^{-1}A_s'A_s(A'A)^{-1}A_j' -$$
$$2\sum_{j=1}^{n}A_j(A'A)^{-1}A_s'A_s(A'A)^{-1}A_s'e_0'(Y_j - A_j\hat{b})' +$$
$$2\sum_{j=1}^{n}A_j(A'A)^{-1}A_s'e_0e_0'A_s(A'A)^{-1}A_j',$$
$$\partial^2[\hat{\Sigma}^{-1}] = -\hat{\Sigma}^{-1}\partial^2[\hat{\Sigma}]\hat{\Sigma}^{-1} + 2\hat{\Sigma}^{-1}\partial[\hat{\Sigma}]\hat{\Sigma}^{-1}\partial[\hat{\Sigma}]\hat{\Sigma}^{-1}.$$

Acknowledgements

Tartu university is gratefully acknowledged for the invitation to the first Tartu seminar on multivariate statistics.

The research was supported by the Swedish Natural Science Research Council.

References

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics* **19** 15-18.

Cook, R.D. (1986). Assessment of local influence (with discussion). *J. R. Statist. Soc. B* **48** 133-169.

Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman & Hall, New York.

De Gruttola, V. D., Ware, J.H. & Louis, T.A. (1987). Influence analysis of generalized least squares estimators. *J. Amer. Statist. Assoc.* **82** 911-917.

Kollo, T. & Neudecker, H. (1991). Asymptotic distribution of eigenvalues and eigenvectors of sample covariance and correlation matrices. Unpublished manuscript.

Magnus, J.R. & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381-394.

Magnus, J.R. & Neudecker, H. (1988). *Matrix Differential calculus with Applications in Statistics and Econometrics.* Wiley, New York.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705-724.

von Rosen, D. (1988). Moments for matrix normal variables. *Statistics* **19** 575-583.

Uppsala University
Department of Mathematics
Thunbergsvägen 3
S-752 38 Uppsala
Sweden

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 94-113.

EXTREMAL MULTIVARIATE DISTRIBUTIONS HAVING GIVEN DISCRETE MARGINALS

Ene-Margit Tiit
University of Tartu, Tartu, ESTONIA

Keywords: *Distributions with given marginals, extremal multivariate distributions, Frechet bounds, minimal and maximal correlations*

1. Set-up of the problem.

Let P_i be given h_i-variate distributions, $i = 1,\ldots,k$; $\sum h_i = m$.

The problem is to describe the set $\Pi = \Pi(P_1, \ldots, P_k)$ of all m-variate distributions in the space $\mathbb{R}^m$, having the given (h_i-variate) marginals P_i, $i=1, \ldots, k$. As the set Π is a convex polyhedra, see Kellerer (1964), the problem can be reduced to the problem of describing the class of all extremal distributions, belonging to Π, and being the vertices of the set Π.

The following question is - how should the extremal distributions be defined? The problem has been solved in an essential way for some special cases.

In the case $h_1 = h_2 = 1$, $k = m = 2$, the classical solution, given by Hoeffding (1940) and Fréchet (1951) is well known. Then two special distributions, the so-called lower and upper Frechet bounds Q^- and Q^+, defined by their distribution functions (d.f.) $H^-(x_1,x_2)$ and $H^+(x_1,x_2)$, are the following:

$$H^-(x_1,x_2) = \max(0, F_1(x_1) + F_2(x_2) - 1), \tag{1}$$

$$H^+(x_1,x_2) = \min(F_1(x_1),F_2(x_2)), \tag{2}$$

where $F_i(x_i))$ denotes the d.f.of the given distribution P_i, i= 1,2.

It is well known (see Fréchet (1951)) that the distributions Q^+ and Q^- have also the maximal and minimal possible correlations r^+ and r^- correspondingly. In this paper we use the following definition:

Definition 1. The k-variate distribution, having all its correlations extremal (minimal or maximal), is said to be extremal.

For the case k > 2 some solutions have also been found.

The maximal distribution, i.e., the distribution, having all maximal correlations, is an immediate generalization of the distribution Q^+, and its d.f. can be constructed with the help of the following formula (see Ruiz-Rivas (1979), Cuadras (1981)).

$$H^+(x_1,\ldots,x_k) = \min_{1\leq i\leq k} F_i(x_i). \tag{3}$$

Ruiz-Rivas (1979) expressed an idea that the number of all extremal k-variate distributions was 2^{k-1}. Tiit (1984) and Rüschendorf (1985) gave the construction of these extremal distributions by their d.f's and in 1986 the author proved that the set of the extremal distributions is a proper subset of the set of all vertices of the set Π, see Tiit (1986).

In the case of distributions P_i having the finite support, also the probability function (p.f.) of the extremal distributions can be constructed, see Frechet (1951), Tiit (1986, 1989).

The purpose of this article is:

- to give an algorithm for the construction of p.f. of arbitrary k-variate extremal distribution with given marginals, having a finite support,
- to study some properties of these extremal distributions,
- to give some ideas for possible practical applications of the concept of the extremal distribution and also the algorithm for its construction.

All constructions are illustrated by computational examples. Also some simulation examples are given.

2.The construction of d.f. of a k-variate extremal distribution.

Let the given marginals P_i be univariate, h_i = 1, i = 1, ..., k. In future we shall use the following notations.

If Q is a k-variate distribution, then $r_{ij}(Q)$ (1 = i,j = k) de-

notes the correlation coefficient between the i-th and j-th component of a random vector $X = (X_1, \ldots, X_k)'$, having the distribution Q.

If P_i and P_j are univariate distributions, then $r^+(P_i,P_j)$ and $r^-(P_i,P_j)$ denote the maximal and minimal possible correlations for all bivariate distributions, having the marginals P_i and P_j. These correlations are defined as $r_{12}(Q^-)$ and $r_{12}(Q^+)$, where the distributions Q^- and Q^+ are constructed by formulae (1) and (2) correspondingly.

Let us denote the set of integers $(1, \ldots, k) = I_k^o$ and let I denote some subset of I_k^o, fulfilling the condition:

$$1 \in I, \tag{4}$$

$I = (i_1, \ldots i_q)$, $1 \leq q \leq k$. I is said to be an index-set.

Let $I^- = I_k^o \setminus I$, $I^- = (j_1, \ldots, j_s)$, where $0 \leq s \leq k-1$, $q + s = k$. It is evident that the number of different subsets I of I_k^o (or pairs (I, I^-)) is equal to 2^{k-1}.

This result is illustrated in Example 1, where all possible index-sets, being contained in $I_4^o = (1,2,3,4)$ are given (see Table 1).

For the case of k-variate distribution Q we can regard the set I as an index-set, defining some q-variate ($q < k$) marginal distribution Q(I) of Q. The distribution Q(I) can also be regarded as the distribution of the q-variate subvector X_I of vector X, $X \sim Q$.

In Tiit (1980) the following theorem is proved:

Theorem 1. Every index-set I defines a k-variate distribution Q_I, belonging to the set Π. The distribution Q_I is defined by its d.f. $H_I(.)$ in the following way:

$$H_I(x_1,\ldots,x_q) = \max(0, \min_{i\in I} F_i(x_i) + \min_{j\in I^-} F_j(x_j) - 1). \tag{5}$$

The distribution Q_I has some important properties which will be listed in the following corollaries.

Corollary 1. In the case $I = I_k^o$ the distribution Q_I is the maximal distribution, defined by the formula (3).

Corollary 2. In general, the distribution Q_I has a q-variate marginal Q^+ and a s-variate marginal Q^- defined by their d.f's $H^+(.)$ and $H^-(.)$ with the help of the formula (3). Q^+ and Q^- are both maximal

distributions in the sets $\Pi(P_{i_1},\ldots,P_{i_q})$ and $\Pi(P_{j_1}, \ldots, P_{j_s})$.

The validity of Corollaries 1 and 2 follows immediately from the formulae (3) and (5).

Corollary 3. Distribution Q_I is an extremal distribution. The values of the correlation coefficients of the distribution Q_I are expressed in the following way:

$$r_{ij}(Q_I) = \begin{cases} r^+(P_i,P_j), \text{ if } (i \in I) \text{ and } (j \in I) \\ \qquad \text{or } (i \in I^-) \text{ and } (j \in I^-), \\ r^-(P_i,P_j), \text{ if } (j \in I) \text{ and } (i \in I^-) \\ \qquad \text{or } (j \in I^-) \text{ and } (i \in I). \end{cases} \tag{6}$$

In Example 1 the structure of all different extremal correlation matrices of the four-variate distributions is demonstrated in special schemes, see Table 2.

Corollary 4. If all the distributions P_i are symmetrical and identical, then every extremal distribution Q_I is concentrated on a line defined by some diagonal of the k-variate unit cube. The maximal distribution is concentrated on the main diagonal.

Corollaries 3 and 4 are proved in Tiit (1986).

3. Some properties of the p.f's of k-variate distributions having a finite support.

Here we regard the p.f. of a k-variate distribution P as a pair (A, p), where A is the support of the distribution P, consisting of finite number ordered points $a_i = (a_1^i, \ldots, a_k^i)$ in space R^k, and p is the vector of probabilities, $p = (p_1, \ldots, p_s)$, $p_i = P(a_i)$, i=1, ..., s.

Let f be some one-to-one transformation $R^k \Longrightarrow R^m$, then $P' = fP$ denotes the m-variate distribution $P' = (fA, p)$ and $P = f^{-1}P'$.

At first let us regard the case, when the distribution P = (A, p) is univariate, and define some special distributions which are important for the following constructions.

Definition 2. The univariate distribution $P_c = (I_s^o, p)$ is said to be canonical.

From here it follows, that as the support of canonical distribution $I_s^o = (1, 2, \ldots, s)$ is uniquely defined by its power s, the canonical distribution is defined by its vector of probabilities $p = (p_1, \ldots, p_s)$.

Definition 3. Let P = (A, p) be a univariate distribution having a finite support of power s. If there exists an increasing transformation c, satisfying the following condition

$$cA = I_s^o \tag{7}$$

then the distribution P is said to be non-decreasing.

In future we shall denote the distribution (cA,p) by cP. As the distribution cP is by Definition 2 canonical, we say that the transformation c is a canonization transformation for distribution P.

As all univariate non-decreasing distributions are equivalent to canonical distributions, we assume in future that every non-decreasing univariate distribution is canonical.

Let us regard now some properties of the support of a multivariate distribution.

Let P = (A, p) be a k-variate distribution having the finite marginal distributions $P_i = (A_i, p_i)$, i = 1, ..., k. Then the following inclusion is true

$$A \subset A_1 * \ldots * A_k,$$

where B * C denotes the Cartesian product of sets B and C.

In the case of finite support A from here the inequality for the power s of A follows:

$$s = s_1 \times \ldots \times s_k, \tag{8}$$

where s_i is the power of the set A_i.

Definition 4. Let P = (A, p) be a k-variate distribution having a finite support, and let its univariate marginals $P_i = (A_i, p_i)$, i = 1, ..., k, be canonical. The distribution P is said to be a completely non-decreasing k-variate distribution in respect of all its marginals if there exists a transformation f,

$$A = f(A_1, \ldots, A_k),$$

fulfilling the following conditions:

- f is non-decreasing in respect of all its arguments,
- f has a unique inverse f^{-1}.

In the special case when k = 2 we say that the distribution is non-decreasing instead of saying that it is completely non-decreasing.

The completely non-decreasing distributions, described by Definition 4, are quite special and rare. In the following we shall define a somewhat milder property for multivariate distributions.

Definition 5. Let P be a k-variate distribution having a finite support. If all bivariate marginal distributions P_{ij} of P (i,j = 1, ..., k, i ≠ j) are non-decreasing, then P is said to be a pairwise non-decreasing distribution.

From the elementary properties of non-decreasing functions the following result can be derived.

Corollary 5. The completely non-decreasing distribution is also non-decreasing pairwise.

From Definition 4 the following important property of the non-decreasing k-variate distributions can be concluded.

Corollary 6. If P is a completely non-decreasing k-variate distribution, then there exists a univariate canonical distribution P', being equivalent with the distribution P in the sense that there exists such transformation c, that P' = cP and P = c^{-1}P'.

Proof. For proving the corollary we construct the transformation c in the following way:

$$c(a_{1j}, \ldots, a_{kj}) = j, \tag{9}$$

where j is the index of the point of the support A of the k-variate distribution P, and at the same time - the value of the point of the support A' of the univariate distribution P'.

Definition 6. Let P be a completely non-decreasing distribution and P'= cP be the univariate canonical distribution, equivalent to P. Then the transformation c is said to be a canonization transformation.

The canonization transformation defined here is a straightforward generalization of the same concept, which was taken into use in the paragraph 3.

4. The construction of p.f. of a bivariate maximal distribution.

Let P_1 and P_2 be two canonical distributions, defined by their p.f's $p_1 = (p_1^1, \ldots, p_s^1)$ and $p_2 = (p_1^2, \ldots, p_t^2)$.

Our aim is to give the algorithm for the construction of the p.f. of the maximal bivariate distribution Q^+, $Q^+(i,j) = q_{ij}$, $i = 1, \ldots, s$, $j = 1, \ldots, t$.

The algorithm, which we shall give, has been described by Frechet (1951), our task is to find the way for generalizing it for the k-variate case, see also Tiit (1986, 1989).

Algorithm 1.

Step 1. $q_{11} := \min(p_1^1, p_1^2)$, $h := 1$, $i_1 := 1$, $j_1 := 1$.

Step h+1 ($h = 1$).

$i := i_h$, $j := j_h$, $h := h + 1$,

if $\sum_{f=1}^{j} q_{if} = p_i^1$, then $q_{ij+1} := \ldots := q_{it} := 0$, $i := i+1$,

if $\sum_{g=1}^{i-1} q_{gj} = p_j^2$, then $q_{ij} := \ldots := q_{sj} := 0$, $j := j+1$,

$$q_{ij} := \min\left(p_i^1 - \sum_{f=1}^{j-1} q_{if},\ p_j^2 - \sum_{g=1}^{i-1} q_{gj}\right),$$

$i_h := i$, $j_h := j$.

If $i_h = s$ and $j_h = t$, then end, $r := h$, else next, h+1-st step.

In Example 2 the usage of Algorithm 1 is illustrated for a pair of given univariate distributions, see Table 3.

As a result, Algorithm 1 gives the construction of bivariate maximal distribution Q^+, including

- the construction of the support B of the distribution Q^+,

$$B = ((i_h, j_h),\ h = 1, \ldots, r),$$

- the construction of the p.f. or the vector of probabilities q,

$$q = (q_h),\ h = 1, \ldots, r.$$

Let us denote the operator, realizing Algorithm 1, by T. Consequently, we have

$$Q^+ = T(P_1, P_2).$$

The operator T and the distribution Q^+ have some important properties. We shall present them in the following paragraph.

5. Some properties of a bivariate maximal distribution.

The first result is concerning the power of the maximal distribution.

Corollary 7. The power r of the support B of the distribution Q^+ satisfies the following inequalities:

$$\max(s,t) \leq r \leq s + t - 1.$$

This fact arises immediately from Algorithm 1, where every step defines one point of the support, and on every step either row index, either column index, either both, increase by one unit. The construction ends when both indices have gained their maximal possible values.

From here it follows, that the power of the support of maximal distribution is, in general, much smaller than the upper bound of the power, given in formula (8).

Corollary 8. The distribution Q^+ is non-decreasing.

Also the proof of this corollary arises immediately from Algorithm 1 - on every step at least one of the indices of the point coordinates (b_i, b_j) increases, consequently, the inequalities

$$i_h \leq i_{h+1}, \quad j_h \leq j_{h+1}$$

hold for h = 1, ..., r-1 and, consequently, the transformation (9) is non-decreasing and also a canonization transformation.

Corollary 9. The operator T is unique.

This assertion follows from the construction, given in Algorithm 1. Here on every step both, the point of support and its probability, were defined uniquely, also the finite number of steps was uniquely determined by the given marginal p.f's.

From Corollary 9 the following important fact can be deduced:

Corollary 10. If a bivariate distribution, belonging to the set $\Pi(P_1, P_2)$ is non-decreasing, then it can be constructed with the help of Algorithm 1.

From Corollaries 7 - 10 the following theorem that fixes one of the most remarkable properties of the maximal bivariate distribution, arises.

Theorem 2. Let P_1 and P_2 be given univariate distributions having the finite supports. The bivariate distribution Q with marginals P_1 and P_2 is maximal if and only if it is nondecreasing.

Proof. From Corollary 8 it follows that the maximal distribution is non-decreasing.

For proving the maximality of a non-decreasing distribution let us suppose that there exists an non-decreasing, but not a maximal distribution W in the set $\Pi(P_1, P_2)$. So as W is non-decreasing, from Corollary 10 it follows that it must be received by the process described in Algorithm 1. But as the construction defines the distribution uniquely, the only possibility is that W = Q.

From Corollaries 7 - 10 it can be concluded that the maximal bivariate distribution is in some sense equivalent with a univariant distribution, its support is an ordered set. This fact can be used for expanding Algorithm 1 for the multivariate case by generalizing the result, expressed in Corollary 4.

6. The construction of the k-variate maximal distribution.

Let us define some special operations for non-decreasing k-variate distributions and their non-decreasing transformations.

Definition 7. Let P = (B, p) be a k-variate distribution and $B \subset R^k$. Let P_1 and P_2 be the g- and h-variate marginals of P, g + h = k. If $B_1 \subset R^h$ and $B_2 \subset R^g$, R^g and R^h are orthogonal, then the marginal distributions P_1 and P_2 are said to be complementary. This fact will be denoted in the following way

$$P = (P_1; P_2).$$

Let P = (B, p) be completely non-decreasing k-variate distribution, having complementary g- and h-variate marginals P_1 and P_2, the canonization transformations of which are denoted by c_1 and c_2 correspondingly. Let us introduce the following transformation of P

$$P' = c_1 P = (c_1 P_1; P_2).$$

Here P' is the h+1-variate non-decreasing distribution $(B'; B_2, p)$, where the support of the initial distribution P has been transformed, but the p.f. has remained unchanged. In the similar way the inverse transformation c_1^{-1} is defined,

$$c^{-1}P' = (c^{-1}P'_1; P_2) = (P_1; P_2),$$

where P'_1 is the canonized univariate distribution, corresponding to

the g-variate distribution P_1. The result of the last transformation is the initial g+h-variate distribution.

In the similar way the transformations $Pc_2 = (P_1, c_2 P_2)$ and $P'c_2^{-1} = (P_1, c_2^{-1} P_2)$, applied to the second marginal P_2, will be used. Also the different transformations c_i and c_j^{-1} can be combined with each other and with other non-decreasing transformations. Notice that all distributions, received as a result of these operations, have the following properties.

1) They are unique.

2) They are non-decreasing,

3) They have the same p.f. In our notation the vector of probabilities is p.

From Theorem 2, Corollaries 5, 7 - 10 and the definition of extremal k-variate distribution the following important result arises.

Theorem 3. If the k-variate distribution Q, having a finite support, is pairwise non-decreasing, then it is also maximal.

Proof. So as every bivariate marginal of the distribution Q is non-decreasing, from Theorem 2 it follows that it is also maximal, and hence all correlation coefficients of the distribution Q are maximal. From the definition it follows that the distribution Q is maximal.

For practical applications the following result, concluded from Theorem 3 and Corollary 6, is useful.

Corollary 11. Every completely non-decreasing k-variate distribution Q, having a finite support, is maximal.

Our next purpose is to give the construction of k-variate maximal distribution. We begin with the case k = 3.

Theorem 4. Let P_1, P_2, P_3 be given univariate canonical distributions, $P_i = (A_i, p_i)$, i = 1,2,3. Then the distribution Q^+,

$$Q^+ = c^{-1} T(cT(P_1, P_2), P_3), \tag{10}$$

where T is the operator, defined by Algorithm 1, and c the canonization transformation, is the maximal distribution in $\Pi(- P_1, P_2, P_3)$.

Proof is constructive. Let us construct the following distributions step by step:

Step 1. With the help of Algorithm 1 construct the maximal distribution by marginals P_1 and P_2:

$$Q = T(P_1, P_2), \quad Q = (B, q), \quad B = ((a_h^1, a_h^2), \ h=1, \ldots, r).$$

Step 2. Canonize the maximal distribution Q and get the equivalent univariate distribution P:

$$P = cQ,\ P = (A,q),\ A = cB = (a_h,\ h = 1,\ \ldots,\ r).$$

Step 3. With the help of Algorithm 1 construct the maximal distribution, having the marginals P (received on the previous step) and P_3 (the third given marginal):

$$Q' = T(P,P_3),\ Q' = (B',q),\ B' = ((a_h,a_h^3),\ h=1,\ \ldots,\ s).$$

Step 4. Use the inverse of canonization transformation for substituting the univariate marginal distribution P by its bivariate equivalent - the maximal distribution Q, having the univariate marginals P_1 and P_2:

$$Q^+ = c^{-1}Q',\ Q^+ = (c^{-1}B',\ q),$$

where

$$c^{-1}B' = ((a_h^1,a_h^2,a_h^3),\ h = 1,\ \ldots,\ s).$$

From the construction the following properties of the distribution Q^+ can be concluded:

1) Q^+ is trivariate, $c^{-1}B'c\ R^3$, as the inverse canonization operation has been applied to the bivariate distribution Q'.

2) Q^+ is completely non-decreasing, as the operation T is non-decreasing in respect to its marginals, and also canonization and its inverse are non-decreasing transformations.

Now the validity of the theorem follows immediately from Corollary 11. The construction, given in this theorem, is illustrated in Example 3.

Theorem 4 can be expanded in the following way.

Theorem 5. Let $P_1,\ \ldots,\ P_k$ be given canonical univariate distributions. Then for h, h = 1, ..., k - 1 the following assertion holds.

Let Q_h be the maximal distribution in $\Pi(P_1,\ \ldots,\ P_h)$.

Then the distribution Q_{h+1}^+,

$$Q_{h+1}^+ = c^{-1}T(cQ_h,\ P_{h+1}),$$

where c denotes the canonization transformation of Q_h, c^{-1} its inverse and T is the operation, given by Algorithm 1, is the maximal distribution in $\Pi(P_1,\ \ldots,\ P_{h+1})$.

Proof is based on mathematical induction and the repeated usage of the construction, given in the proof of Theorem 4.

Theorem 5 is illustrated in the Example 4, where the four-variate maximal distribution is built and also analyzed.

The Theorem 4 has also another expansion, which is useful for us in the future. As its derivation is similar to that one of Theorem 5 (the only difference is that in this case both arguments, used in Algorithm 1, have been received by canonization of multivariate non-decreasing distributions), we omit it and present the result in the following Corollary 12.

Corollary 12. Let P_1, ..., P_k be given canonical univariate distributions, Q_h and Q_g the maximal distributions in the sets $\Pi(P_1, \ldots, P_h)$ and $\Pi(P_{h+1}, \ldots, P_{h+g})$ correspondingly, $h + g = k$.

Then distribution Q^+_{h+g}, defined by the formula

$$Q^+_{h+g} = c_1^{-1}(T(c_1Q_h, c_2Q_g))c_2^{-1},$$

where c_1 and c_2 denote the canonization transformations of h- and g-variate distributions Q_h, Q_g and c_1^{-1}, c_2^{-1} are their inverses, is maximal in $\Pi(P_1, \ldots, P_{h+g})$.

7. The construction of arbitrary k-variate extremal distribution.

Our first aim is to find the construction of minimal bivariate distribution. For this purpose we shall define the operation of inversion of a distribution and also the concept of a non-increasing distribution.

Definition 8. Let $P = (A, p)$ be a k-variate distribution having a finite support, $A = (a_1, \ldots, a_s)$, $p = (p_1, \ldots, p_s)$. The distribution $-P = (-A, -p)$, where $-A = (a_s, \ldots, a_1)$ and $-p = (p_s, \ldots, p_1)$ is said to be the inverse distribution P.

Here we must mention that the inversion does not change the distribution in the probabilistic sense. This transformation is needed only for some special constructions.

Definition 9. Let P be a k-variate non-decreasing distribution Then the distribution $-P = (-A, -p)$ is said to be non-increasing.

With the help of the concept of inversion of the distributions the following assertion can be concluded.

Theorem 6. Let P_1 and P_2 be canonical univariate distributions. Then the bivariate distribution $Q^- = T(-P_1, P_2)$ is the minimal distribution.

The validity of the assertion is known already by the construction, given by Frechet (1951) and is also used in Tiit (1986, 1989). We illustrate the Theorem in Example 5.

It must be stressed that this construction is not unique, as it follows from the next corollary.

Corollary 13. Let P_1 and P_2 be canonical univariate distributions. Then the bivariate distribution $Q^* = T(P_1,-P_2)$ is also the minimal distribution.

The connection between the distributions Q^- and Q^* follows from the evident equality

$$-(-P) = P,$$

consequently, we have $Q^* = T(-P_1,-(-P_2)) = -Q^-$. As the inversion does not influence the probabilistic nature of the distribution, it can be inferred that the minimal distribution is unique, but its p.f. can be written in two different ways.

Therefore, in the bivariate case the properties of minimal distribution are dual to these of maximal ones, e.g. the minimal distribution is non-increasing, unique, and equivalent to a univariate distribution.

But for the multivariate case the assertion is not valid. In spite of that, the fact, expressed in Theorem 6, can be used for giving the construction of arbitrary extremal distribution.

Let $P = (P_1; P_2)$ be completely non-decreasing k-variate distribution, P_1 and P_2 being complementary marginals of P and having correspondingly the canonization transformations c_1 and c_2. Consequently, $P'_1 = c_1P_1$ and $P'_2 = c_2P_2$ are univariate canonical distributions.

Let us use the operation T for the pair $-P'_1$ and P'_2, and by Theorem 6 we get as a result the non-increasing bivariate distribution $Q' = (B', q), = (-P'_1; P'_2)$. Using the inverse canonization transformations c_1^{-1} and c_2^{-1} for both marginals, we receive the k-variate distribution

$$Q = (Q_1;Q_2) = (-c_1^{-1}P_1; c_2^{-1}P_2),$$

having the following properties:

- both marginal distributions Q_1 and Q_2 are completely non-decreasing and hence maximal,

- all bivariate marginals P_{ij}, where P_i is a marginal of Q_1 and P_j is a marginal of Q_2, are non-increasing.

From the definition of the extremal distribution it follows that the distribution Q is extremal. Analogically, a more general result can be proved.

Theorem 7. Let P_1, ..., P_k be given univariate marginal distributions, and let I be an index-set, fulfilling the condition (4). Let the distributions P_I and P_{I^-} be the maximal distributions, constructed by Theorem 5. Then the distribution

$$P_{I,I^-} = c_1^{-1}(T(c_1 P_I, -c_2 P_{I^-}))c_2^{-1}$$

is the extremal distribution, fulfilling the conditions (6).

Proof of the theorem is based on the construction that is similar to that one used in Theorem 4.

The illustration of Theorem 7 is given in Example 4.

8. Some applications.

1. The concept of extremal correlations.

In the practice of data analysis the problem of the estimation of correlations occurs rather often, and it is usual to compare the values of correlation coefficients between some variables with the values -1 and +1 as minimal and maximal possible values.

But as in the case of empirical data usually the minimal and maximal values can never be reached, in the estimation of the dependencies between the variables it would be useful to compare the empirical correlations with the minimal and maximal correlation coefficients for the given marginal (empirical) distributions.

2. The extremal distributions.

Often for some population to be examined, the marginal distributions are well known from previous studies, but the multivariate distribution is not known. For designing the survey study the assumption that the variables to be studied are independent would usually be made. In some cases the calculation of one 'least favourable' multivariate distribution should be more effective. As it is in fact one of the extremal distributions, it can be calculated by the scheme, given in this paper. Sometimes it seems to be rational to calculate some set of possible extremal distributions of the most important characteristics of the population and to use this information in survey planning.

9. Examples.

Example 1. Let us regard the set of all index-sets, belonging to the set $I^o = (1, 2, 3, 4)$ and fulfilling the condition (4), see Table 1.

Table 1.

$I^o = (1, 2, 3, 4)$,	$I_4 = (1, 2)$,
$I_1 = (1, 2, 3)$,	$I_5 = (1, 3)$,
$I_2 = (1, 2, 4)$,	$I_6 = (1, 4)$,
$I_3 = (1, 3, 4)$,	$I_7 = (1)$.

In the following we give the correlation matrices, corresponding to all index-sets in a schematical way. From the matrices only the upper triangles are presented, and the maximal and minimal correlation coefficients are denoted by signs + and - correspondingly, see the following table.

Table 2.

I^o	I_1	I_2	I_3
+ + +	+ + -	+ - +	- + +
+ +	+ -	- +	- -
+	-	-	+

I_4	I_5	I_6	I_7
+ - -	- + -	- - +	- - -
- -	- +	+ -	+ +
+	-	-	+

Two of these extremal distributions, corresponding to I^o and I_4, will be studied in detail in Examples 4 and 5.

Example 2. Let us regard two canonical distributions, characterized by their p.f's:

$$p_1 = (0.4, 0.3, 0.3), \quad p_2 = (0.1, 0.2, 0.3, 0.4).$$

Our aim is to calculate the maximal distribution $Q^+ = Q(1,2)$, using Algorithm 1. Technically it is convenient to write the common distribution step by step into the usual bivariate table, using the given marginals that are written in the last column and last row of Table 3.

Table 3.

	1	2	3	4	p_1
1	0.1	0.2	0.1	0	0.4
2	0	0	0.2	0.1	0.3
3	0	0	0	0.3	0.3
p_2	0.1	0.2	0.3	0.4	1.0

By Algorithm 1 we get on the first step $q_{11} = \min(0.1, 0.4) = 0.1$, $i_1 = j_1 = 1$; on the second step we see that $p_1^2 = q_{11} = 0.1$, hence the remaining probabilities in the first column equal zero, and $j_2 = j_1 + 1 = 2$. As $\min(0.4 - 0.1, 0.2) = 0.2$, we get $q_{12} = 0.2$, etc.

The correctness of the construction can be checked by summarizing the probabilities of common distribution by rows and columns and comparing them with the given marginals.

Example 3. Let us regard three canonical distributions, characterized by their p.f's:

$p_1 = (0.4,0.3,0.3)$, $p_2 = (0.1,0.2,0.3,0.4)$, $p_3 = (0.2,0.6,0.2)$.

Our aim is to calculate the maximal distribution $Q^+ = Q(1,2,3)$. Here we can use the results, received in Example 2, where the two first distributions were used. The support $B^+(1,2)$ of the distribution $Q^+(1,2)$ is the following:

$$B^+(1,2) = ((1,1), (1,2), (1,3), (2,3), (2,4), (3,4)),$$

and the p.f. is

$$q^+(1,2) = (0.1, 0.2, 0.1, 0.2, 0.1, 0.3).$$

Here we stress that the ordering in both vectors B^+ and q^+ is the same.

Now let us canonize the distribution Q^+, see Table 4,

Table 4.

cb_i	1	2	3	4	5	6
b_i	(1,1),	(1,2),	(1,3),	(2,3),	(2,4),	(3,4)

and as a result we get the univariate canonical distribution $Q'(1,2)$, having the support $B' = (1, 2, 3, 4, 5, 6)$ and the same p.f. as $Q^+(1,2)$.

The following step, which we must make, is the usage of the transformation T for the pair of univariate canonical distributions $Q'(1,2)$ and P_3 and get the maximal distribution $Q^{++} = Q^+(Q',P_3)$, see Table 5.

Table 5.

	1	2	3	4	5	6	p_3
1	0.1	0.1	0	0	0	0	0.2
2	0	0.1	0.1	0.2	0.1	0.1	0.6
3	0	0	0	0	0	0.2	0.2
$q(1,2)$	0.1	0.1	0.1	0.2	0.1	0.3	1.0

For getting the trivariate maximal distribution $Q^+(1,2,3)$, we must use the inverse canonization transformation for the first marginal of the bivariate distribution Q^{++}.

The support B^{++} of the distribution Q^{++} is the following:

$$B(1,2,3)=((1,1), (2,1), (2,2), (3,2), (4,2), (5,2), (6,2), (6,3))$$

and the p.f. is

$$q(1,2,3) = (0.1, 0.1, 0.1, 0.1, 0.2, 0.1, 0.1, 0.2).$$

For the calculation of the support of the trivariate distribution, we write the transformation c^{-1} with the help of Table 4, where in the first row the values of the support of the canonized distribution B', and in the second row - the original values of $B = c^{-1}B'$, are given.

Now let us substitute the values of the points $c^{-1}b_i$ in the support of the distribution Q^{++}. As a result we get the following eight points of R^3, see Table 6.

Table 6.

1	2	3	4	5	6	7	8
(1,1,1)	(1,2,1)	(1,2,2)	(1,3,2)	(2,3,2)	(2,4,2)	(3,4,2)	(3,4,3)

The result of the calculations is the maximal three-variate distribution $Q^+ = Q(1,2,3)$ that can be written in the following table:

Table 7.

point index	X_1	X_2	X_3	probability
1	1	1	1	0.1
2	1	2	1	0.1
3	1	2	2	0.1
4	1	3	2	0.1
5	2	3	2	0.2
6	2	4	2	0.1
7	3	4	2	0.1
8	3	4	3	0.2

From here it is simple to see that all the marginals are non-decreasing in respect to each other and also the distribution Q^{++} is completely non-decreasing. Besides that, with the help of Table 6 it is easy to check that all the marginals of the distribution Q^{++} have the given marginals.

Example 4. Let us regard four canonical distributions, characterized by their p.f's:

$$p_1 = (0.4,0.3,0.3), \quad p_2 = (0.1,0.2,0.3,0.4),$$
$$p_3 = (0.2,0.6,0.2), \quad p_4 = (0.1,0.1,0.5,0.3).$$

Our aim is to calculate the maximal distribution $Q^+(1,2,3,4)$. Here we can use the results, received in the Examples 2 and 3, where the three first distributions were used.

Now we must use the operation T for the canonized distribution $B'(1,2,3)$, having the eight-values support, and the fourth given marginal P_4, see Table 8.

Table 8.

	1	2	3	4	5	6	7	8	p_4
1	0.1	0	0	0	0	0	0	0	0.1
2	0	0.1	0	0	0	0	0	0	0.1
3	0	0	0.1	0.1	0.2	0.1	0	0	0.5
4	0	0	0	0	0	0	0.1	0.2	0.3
	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	1.0

The result - the maximal four-variate distribution - can also be written in the following table:

Table 9.

point index	X_1	X_2	X_3	X_4	probability
1	1	1	1	1	0.1
2	1	2	1	2	0.1
3	1	2	2	3	0.1
4	1	3	2	3	0.1
5	2	3	2	3	0.2
6	2	4	2	3	0.1
7	3	4	2	4	0.1
8	3	4	3	4	0.2

From here we see that the support of the maximal 4-variate distribution consists of the 8 points only, while the set $A_1 * A_2 * A_3 * A_4$ consists of 3 * 4 * 3 * 4 = 144 points.

For the calculation of the all maximal correlations the pairwise maximal bivariate distributions can be used. For the distribution P_{12} we use Table 3, and the standard calculation of correlation coefficient gives $r^{\dagger}_{12} = 0.842701$. For the following pairs we construct the corresponding maximal bivariate distributions, see the following tables.

Table 10.

P_{13}

	1	2	3
1	0.2	0.2	0
2	0	0.3	0
3	0	0.1	0.2

$r_{13} = 0.761387$

Table 11.

P_{14}

	1	2	3	4
1	0.1	0.1	0.2	0
2	0	0	0.3	0
3	0	0	0	0.3

$r_{14} = 0.807573$

Table 12.

P_{23}

	1	2	3
1	0.1	0	0
2	0.1	0.1	0
3	0	0.3	0
4	0	0.2	0.2

$r_{23} = 0.790569$

Table 13.

P_{24}

	1	2	3	4
1	0.1	0	0	0
2	0	0.1	0.1	0
3	0	0	0.3	0
4	0	0	0.1	0.3

$r_{24} = 0.894427$

Table 14.

P_{34}

	1	2	3	4
1	0.1	0.1	0	0
2	0	0	0.5	0.1
3	0	0	0	0.2

$r_{34} = 0.883883$

A simulation study (1000 points) gave the following result for the maximal correlation matrix:

Table 15.

	X_2	X_3	X_4
X_1	0.8516	0.7728	0.8173
X_2		0.7954	0.8992
X_3			0.8846

Example 5. Let us regard the same given univariate distributions as in Example 4, and let us use the index-sets, introduced in Example 1. Let us assume that the index-set $I_4 = (1, 2)$ is given. As it follows, we must construct two maximal distributions $Q^+(1,2)$ and $Q^+(3,4)$, then to convert one of them and after that once more the operation T must be used.

As the distributions $Q^+(1,2)$ and $Q^+(3,4)$ are already found in Tables 3 and 14, then the next step which we must make, is to convert one of them and then use Algorithm 1.

In the Table 16 the support of $Q^+(3,4)$, its canonized support, the p.f., inverse canonized support and inverse p.f. are given.

Table 16.

(b_i,b_j)	h	p_h	$-B_c$	$-p_h$
(1,1)	1	0.1	5	0.2
(1,2)	2	0.1	4	0.1
(2,3)	3	0.5	3	0.5
(2,4)	4	0.1	2	0.1
(3,4)	5	0.2	1	0.1

We see that for getting the inversed distribution, the indices of the support must be read from the bottom of the table.

Now the following step is to use the operation T for the distributions $cQ^+(1,2)$ and $-cQ^+(3,4)$. This operation will be illustrated in Table 17.

Table 17.

	5	4	3	2	1	$q^+(12)$
1	0.1	0	0	0	0	0.1
2	0.1	0.1	0	0	0	0.2
3	0	0	0.1	0	0	0.1
4	0	0	0.2	0	0	0.2
5	0	0	0.1	0	0	0.1
6	0	0	0.1	0.1	0.1	0.3
$-q^+(3,4)$	0.2	0.1	0.5	0.1	0.1	

The multivariate distribution, received as the result of this step, can be written in the form of Table 18.

Table 18.

Point index	X_1	X_2	X_3	X_4	probability
1	1	1	3	4	0.1
2	1	2	3	4	0.1
3	1	2	2	4	0.1
4	1	3	2	3	0.1
5	2	3	2	3	0.2
6	2	4	2	3	0.1
7	3	4	2	3	0.1
8	3	4	1	2	0.1
9	3	4	1	1	0.1

From Table 18 we see that the bivariate distributions P(1,3), P(1,4), P(2,3) and P(2,4) are non-increasing. We give also the theoretical values of these minimal correlations.

$$r(1,3) = -0.765510,\ r(1,4) = -0.807573,$$
$$r(2,3) = -0.779057,\ r(2,4) = -0.782624.$$

Here we can see that in the case, when at least on of the univariate marginals is symmetrical, the equality

$$r_{ij}^- = -r_{ij}^+ \qquad (11)$$

holds. But we must mention that the symmetricity is not a necessary condition for the equality (11). The pair of univariate marginals P_2 and P_4 demonstrates that there exist considerable differencies between the absolute values of the extremal correlations.

The simulation experiment (number of points 2000) gave the following values of the correlation matrix, see Table 19.

Table 19.

	X_2	X_3	X_4
X_1	0.8382	-0.7565	-0.7967
X_2		-0.7383	-0.7692
X_3			0.8889

In both simulation studies the fitting between the theoretical and simulated data is quite satisfactory.

References

Cuadras, C.M.(1981) Methods de analisis multivariante. Barcelona, p.52.

Frechet, M. (1951) Sur les tableaux de correlation dont les marges sont donnees. - Ann. Univ. Lyon. Sect. A, fasc. 14, p. 53 - 77.

Hoeffding, W. Masstabinvariante Korrelationstheorie. - Schr. Math. Inst. u. Inst. Angew. Math. Univ. Berlin, 5, S. 179 - 233.

Kellerer, J.C. (1964) Masstheoretische Marginalprobleme. - Math. Ann., 153, H. 3, S. 168 - 198.

Ruiz-Rivas, C. (1979) Contribuciones a la construccion de la distribuciones multivariantes co marginales fijas. Tesis Doctoral, Fac. Mat, Univ. Barcelona, (manuscript).

Rüschendorf, L. (1985) Construction of multivariate distributions with given marginals. Ann. Inst. Stat. Math, 37, p. 225 - 233.

Tiit, E.-M. (1984) Definition of random vectors with given marginal distributions and given correlation matrix. - Acta et Commentationes Universitatis Tartuensis, 685, p. 21 - 36.

Tiit, E. (1986) Random vectors with given arbitrary marginals and given correlation matrix. - Acta et Commentationes Universitatis Tartuensis, 773, p 14 - 38.

Tiit, E.-M. (1989) Generator of multivariate random numbers having given marginal distributions and correlation matrix. - Applications of Multivariate Statistical Analysis in Economics and Quality Control. Proceedings, I, Tartu, p.97 (in Russian).

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 114-120.

EMPIRICAL INVESTIGATION OF THE ESTIMATES OF PROBABILITY

Liina-Mai Tooding
University of Tartu, Tartu, ESTONIA

Keywords: *Multinomial distribution, effect of the sample size, confidence limits*

1. Introduction. The purpose of this empirical work is to analyse the effect of sample size on the values of the estimates of the parameter $\vec{p} = (p_1, \ldots, p_k)$ of the multinomial distribution. The treatment is based on the set of real data. Changes in the empirical distribution of the variable are investigated when the sample size decreases systematically. On the each sample size N a subset of the initial data set is randomly selected. It appears that the value of the estimator calculated from the subset is quite near to this one found from the initial data set. It also holds for the samples which are traditionally considered as insufficiently small.

2. Set-up of the problem. The most frequent questions put by a client to a statistician in the applied fields of mathematical statistics are: how many observations we need, how large must the sample size be in the task. This problem is quite important from the economical point of view but so far, unfortunately, a solution, suitable for handling in practice, has not been found. It holds, in particular, in the case of the data sets derived from researches in social sciences where, as a rule, nominal scales are used and the population

distributions is supposed to be multinomial. Let us recall the definition of this theoretical distribution.

Let $E_1,\ldots,E_k$ be a system of events excluding one another and let the probability of event E_i be equal to p_i, $\sum_{i=1}^{k} p_i = 1$. Let us have a sequence of independent cases, in each of which just one of these k events may be observed. Then the joint distribution of the random variables $n_1,\ldots,n_k$, representing the numbers of occurrences of the events $E_1,\ldots,E_k$ respectively in N cases, is defined by multinomial distribution:

$$P(n_1,\ldots,n_k) = N!\prod_{i=1}^{k} p_i^{n_i}/n_i!,$$

where $\sum_{i=1}^{k} n_i = N$, $n_i \geq 0$.

The multinomial distribution is specified by parameters N, k and $\vec{p} = (p_1,\ldots,p_k)$. In an applied task the parameters $p_1,\ldots,p_k$ are interpretable by contents (as the probabilities of distinct values on the scale of the variable), the parameters N and k are previously fixed. The purpose is to derive estimates $\hat{p}_i$ of p_i, $i = 1,\ldots,k$ as precisely as possible. At the same time the estimates are depending on the parameter N: $\hat{p}_i = \hat{p}_i(N)$. The aim of the present work is to investigate the changes in values $(\hat{p}_1(N),\ldots,\hat{p}_k(N))$ by varying sample size N and the number of values k.

3. The criterion of confidence limits. Quite often practical investigator determines the needed sample size by means of predetermining the width of statistical confidence limits for the estimator of probability. Commonly the case of binomial distribution (k=2) is considered (Walker, Lew (1953) for example). Let us illustrate this approach in the case of multinomial distribution in order to give a background for the characterization of the empirical behaviour of estimates $\hat{p}_i(N)$.

As it is well known, the relative frequencies are the maximum likelihood estimators of the parameters of multinomial distribution:

$$\hat{p}_i = \frac{n_i}{N},$$

where n_i is the number of occurrences of the event E_i in the sample. Then according to Johnson, Kotz (1969), ch. 11.2, asymptotic confidence intervals for $p_1,\ldots,p_k$ with the significance level α are the following

$$\frac{\chi^2_{k-1,1-\alpha}+2n_i \pm \chi^2_{k-1,1-\alpha}\left(\chi^2_{k-1,1-\alpha}+4n_i(N-n_i)/N\right)^{1/2}}{2(N+\chi^2_{k-1,1-\alpha})}, \qquad (1)$$

where $\chi^2_{k-1,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with $k-1$ degrees of freedom. The limits depend on the value of the parameter: for $p_i=0.5$ the confidence interval is relatively wider.

If in an applied task the desired width of confidence interval is predetermined, then by (1) we can find the minimal sample size which guarantees the given statistical reliability of the estimators of probability.

Table 1.

Dependence of the width of 95% confidence limits on sample size.

(width x 100)

	p=1/16			p=1/4			p=1/2		
N	k=2	k=4	k=16	k=2	k=4	k=16	k=2	k=4	k=16
2000	2.1	3.0	5.5	3.8	5.4	9.6	4.4	6.2	11.1
1800	2.2	3.2	5.8	4.0	5.7	10.2	4.6	6.6	11.7
1600	2.4	3.4	6.2	4.2	6.0	10.8	4.9	7.0	12.4
1400	2.5	3.6	6.6	4.5	6.5	11.5	5.2	7.4	13.2
1200	2.7	3.9	7.1	4.9	7.0	12.4	5.6	8.0	14.3
1000	3.0	4.3	7.9	5.3	7.6	13.6	6.2	8.8	15.6
800	3.4	4.8	8.8	6.0	8.5	15.2	6.9	9.8	17.4
600	3.9	5.6	10.3	6.9	9.8	17.4	7.9	11.3	20.0
400	4.8	6.9	12.8	8.4	12.0	21.2	9.7	13.8	24.3
200	6.8	9.9	18.8	11.9	16.9	29.4	13.7	19.4	33.3

In Table 1 some examples of the confidence interval for $\hat{p}_i - \hat{p}_i(N)$ at the significance level 0,05 are given. The sample size decreases from 2000 by a step of 200 until 200. At that the multinomial distribution models by k=2, k=4 and k=16 are considered and the calculations for the cases $p_i = 1/2$, $p_i = 1/4$ and $p_i = 1/16$ are accomplished. We can interpret these cases as the situation where events on the scale of the variable have equal probabilities of occurrence.

A moderate growth of the width of the confidence interval appears at N=600, but a remarkable steeper growth begins at N=400. In the table the continuous line encircles the region where the width of the confidence interval does not exceed 5% and the broken line limits a 10%-region in our example.

4. Experimental survey of the sample size. Let us now consider experimentally the changes in the values of estimates $\hat{p}_i - \hat{p}_i(N)$ when sample size N varies. We apply the empirical distributions of variables which are obtained from the real stream of statistical data sets. The variables have the 4- or 16-value scales: k=4 and k=16. The total number of investigated probabilities p_i is equal to 764 in the case k=4 and 880 in the case k=16.

The basis of our experiment consists of the set of estimates of the parameter $\hat{p}_i$ calculated from subsamples of the initial sample. The sample size of subsamples varies from N=2037 (sample size of initial sample) by a step of 0,1N until to the value 0,1N.

In order to characterize the estimates $\hat{p}_i(N_j)$, $j=1,\ldots,9$, $N_j=0,1 \cdot j \cdot N$ we use the differences

$$d_i(j) = |\hat{p}_i(N_j) - \hat{p}_i(N)| \qquad (2)$$

Therefore the estimates obtained on the basis of the initial sample are considered as "true" values since the real values of parameters in the empirical approach are unknown.

The experimental basis of the present work consists of empirical distributions of indicators which are randomly selected from the real data sets of sociological research work.

The estimated probability represents quite a typical set of distributions which occur in this area.

In Table 2 some parameters of the distribution of estimates $\hat{p}_i(N)$ for sample size N=2037 are represented.

Table 2.

Distribution of empirical probabilities

k	minimum	lower quartile	median	upper quartile	maximum
4	0	0.01	0.06	0.40	1.00
16	0	0.05	0.03	0.10	0.40

It appears that the small probabilities dominate.

The calculations in the present experiment were mainly accomplished by means of the statistical package "Statgraphics" 3.0.

5. **Variability of estimates.** If we compare the empirical distributions of probability estimates calculated from the subsamples with different sizes with these calculated from the original sample, then we can detect only a few significant differences between them. In Table 3 the results of the comparation by means of the Kolmogorov-Smirnov test are represented.

Table 3.

Comparison of distributions at N and N_1.

k	n	Kolmogorov-Smirnov test statistic Approximate significance level					
		N_8	N_4	N_5	N_3	N_2	N_1
4		0.024	0.038	0.03	-	0.094	0.114
	764	1.0	1.0	1.0		0.002	0.0
16		0.027	0.027	0.034	0.048	0.095	0.162
	880	1.0	1.0	1.0	0.30	0.001	0.0

On the significance level 0,005 the distributions differ one from another for N_1 and N_2. The absence of differences between total distributions does not exclude the possibility of individual differences between the probability estimates but it still indicates the fact that an accordance between estimates exists in general lines.

In order to investigate the individual differences, the absolute derivations $d_i(i)$ according to (2) were used. In Table 4 the variability of derivations concerning the values of the

Table 4.

Variability of absolute deviations ($x10^2$).

Limits of probabilities	number of estimates	N_9 max	N_8 max	N_5 max	N_3 max	N_2 max	N_1 max
k=4							
0...0,0625	382	0,3	0,5	1,3	1,6	2,3	3,1
0,0625...0,25	137	0,6	1,1	2,5	4,0	4,2	5,8
0,25...0,50	78	1,0	1,1	3,4	6,1	5,1	6,3
0,50...0,75	53	0,9	1,0	2,4	4,8	4,5	5,3
0,75...0,9375	64	0,5	0,8	2,1	3,4	3,7	3,2
0,9375...1,0	50	0,3	0,5	1,0	1,2	2,0	1,3
k=16							
0...0,0625	582	0,5	0,6	1,2	2,0	3,6	3,7
0,0625...0,25	265	0,9	1,1	2,9	5,0	4,2	6,4
0,25...0,50	33	1,1	1,0	2,8	3,9	3,3	5,9

estimates and the level of the sample size is shown. In each case the maximal derivation is given.

It appears that deviations from the initial estimates of $\vec{p}$ are surprisingly small. Therefore in an practical task a more modest sample size than commonly practiced may sometimes be used. The increase of the median of the deviation begins at sample sizes 400-600 as it could also be seen from the confidence limit criterion. It is notable that

the maximum deviance is only 0.06 in the both cases k=4 and k=16.

The empirical investigation of robustness of statistical estimates with the help of the approach given above is quite labour-consuming, in particular, in the case of multinomial models. In spite of that such an empirical "look at-the-data" approach may be perspective in certain limits and it allows us to be convinced that there are favourable statistical qualities of estimates for practical applications which could not be deduced from the strict theoretical assumptions.

References.

Johnson, N.L., Kotz, S. (1969). Distributions in Statistics Discrete Distributions. Houghton Mifflin Comp., Boston.

Walker, H.M, Lew, J. (1953). Statistical Inference. Holt, Rinehort and Winston, New York.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 121-131.

EDGEWORTH EXPANSION FOR MULTIPLE CORRELATION COEFFICIENT UNDER AN ELLIPTICAL POPULATION

Imbi Traat
University of Tartu, Tartu, ESTONIA

Keywords: *Elliptical distribution, cumulants, multiple correlation coefficient, Edgeworth expansion*

1. Introduction

Let x be (p+1)-dimensional elliptically distributed random vector, see Fang, Kotz, NG (1990):

$$x \sim EC_{p+1}(0, V, \psi). \qquad (1)$$

That means that the characteristic function of x has the form

$$\phi(t) = \psi(t^T V t)$$

for some scalar function ψ which is called a characteristic generator. The distribution of x is symmetric with

$$Ex = 0, \quad cov(x) = \Sigma = [\kappa_{ij}] = k_2 V, \quad k_2 = -2\psi'(0).$$

Berkane and Bentler (1986) have found that higher moments of x depend only on its second moments κ_{ij} and on the derivatives of characteristic generator:

$$\mu_{i_1,i_2,\dots i_{2m}} = c(m)\sum \kappa_{i_1,i_2}\kappa_{i_3,i_4}\cdots\kappa_{i_{2m-1},i_{2m}},$$

where

$$c(m) = \frac{\psi^{(m)}(0)}{[\psi'(0)]^m}$$

and the summation is over all the $(2m)!/(2^m m!)$ possible groupings of the subscripts into pairs. Actually, the sum above presents the 2m-order moment of $N(0,\Sigma)$. So all the moments of arbitrary elliptical distribution are proportional to the moments of corresponding order of normal distribution. Due to Traat (1990) the higher cumulants of elliptical distribution are proportional in the same way:

$$\kappa_{ijkl} = k_4\sum^{3}\kappa_{ij}\kappa_{kl} = k_4(\kappa_{ij}\kappa_{kl} + \kappa_{ik}\kappa_{jl} + \kappa_{il}\kappa_{jk}), \tag{2}$$

$$\kappa_{ijklab} = k_6\sum^{15}\kappa_{ij}\kappa_{kl}\kappa_{ab}, \tag{3}$$

$$\kappa_{ijklabcd} = k_8\sum^{105}\kappa_{ij}\kappa_{kl}\kappa_{ab}\kappa_{cd}, \tag{4}$$

where k_4, k_6, k_8 are constants expressed through derivatives of characteristic generator.

$$k_4 = \psi''(0)/\psi'^2(0) - 1, \tag{5}$$

$$k_6 = [\psi'''(0) - 3\psi''(0)\psi'(0)]/\psi'^3(0) + 2. \tag{6}$$

$$k_8 = [\psi^{IV}(0) - 4\psi'''(0)\psi'(0) - 3\psi''^2(0) + 12\psi''(0)\psi'^2(0)]/\psi'^4(0) - 6. \tag{7}$$

Let x be partitioned as

$$x = (x_0 \vdots x^{*T}), \quad \text{where } x^{*T} = (x_1, \dots, x_p),$$

causing the following structure for the covariance matrix:

$$\Sigma = [\sigma_{ij}] = \begin{bmatrix} \sigma_{00} & \vdots & \sigma_0^T \\ \dots & \vdots & \dots \\ \sigma_0 & \vdots & \Sigma_* \end{bmatrix},$$

where

$$\sigma_{00} = Ex_0^2, \quad \sigma_0 = Ex_0x^*, \quad \Sigma_* = Ex^*x^{*T}.$$

The square of multiple correlation coefficient between

x_0 and x^* can be expressed as

$$\rho^2 = \frac{\sigma_0^T \Sigma_*^{-1} \sigma_0}{\sigma_{00}} .$$

Let

$$x_1, x_2, \ldots, x_n$$

be the sample with sample covariance matrix

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

and with sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Consider S to be partitioned in the same way as Σ:

$$S = [s_{ij}] = \begin{bmatrix} s_{00} & \vdots & s_0^T \\ \cdots & \vdots & \cdots \\ s_0 & \vdots & S_* \end{bmatrix}$$

Then the sample estimate of ρ^2 is expressed as

$$r^2 = \frac{s_0^T S_*^{-1} s_0}{s_{00}} .$$

Our aim is to find the first four cumulants of r^2 and to present the Edgeworth expansion up to the order 1/n for the statistic

$$T_n = \sqrt{n}(r^2 - \rho^2) .$$

2. The Edgeworth expansion of T_n

It is known that under some conditions, see Bhattacharya and Ghosh (1978), the Edgeworth expansion of several statistics is valid with certain accuracy. The expansion can be expressed in the form, see e.g. Traat (1986):

$$P(T_n < x) = \Phi(x) - \frac{1}{\sqrt{n}} [\gamma_1 \Phi'(x) + \frac{1}{6}\gamma_3 \Phi'''(x)] +$$

$$+ \frac{1}{n} [\frac{1}{2}(\gamma_2 + \gamma_1^2)\Phi''(x) + \frac{1}{24}(\gamma_4 + 4\gamma_1\gamma_3)\Phi^{IV}(x) +$$

$$+ \frac{1}{72}\gamma_3^2 \Phi^{VI}(x)] + o(n^{-1}) \qquad (8)$$

where

$\Phi(x)$ is the c.d.f. of $N(0, \gamma_0)$, $\Phi'(x)$, ..., $\Phi^{VI}(x)$ are the derivatives of $\Phi(x)$ expressed by the Hermite polynomials and density function $\varphi(x)$ of $N(0, \gamma_0)$ as

$$\Phi'(x) = \varphi(x) ,$$

$$\Phi''(x) = -\frac{x}{\gamma_0^2}\varphi(x) ,$$

$$\Phi'''(x) = \frac{1}{\gamma_0^2}\left(\frac{x^2}{\gamma_0^2} - 1\right)\varphi(x) ,$$

$$\Phi^{IV}(x) = -\frac{1}{\gamma_0^4}\left(\frac{x^3}{\gamma_0^2} - 3x\right)\varphi(x) ,$$

$$\Phi^{V}(x) = \frac{1}{\gamma_0^4}\left(\frac{x^4}{\gamma_0^4} - \frac{6x^2}{\gamma_0^2} + 3\right)\varphi(x) ,$$

$$\Phi^{VI}(x) = -\frac{1}{\gamma_0^6}\left(\frac{x^5}{\gamma_0^4} - \frac{10x^3}{\gamma_0^2} + 15x\right)\varphi(x) ,$$

$\gamma_0, \gamma_1, \ldots, \gamma_4$ are the leading terms of the cumulants of statistic T_n .

In this paper our task is to find the expressions of the leading terms of the first four cumulants of statistic

$$T_n = \sqrt{n}(r^2 - \rho^2)$$

under elliptical population. They appear in expansions of cumulants $\kappa_i(T_n)$ as follows:

$$\kappa_1(T_n) = \frac{1}{\sqrt{n}}\gamma_1 + O(n^{-3/2}) ,$$

$$\kappa_2(T_n) = \gamma_0 + \frac{1}{n}\gamma_2 + O(n^{-2}) ,$$

$$\kappa_3(T_n) = \frac{1}{\sqrt{n}}\gamma_3 + O(n^{-3/2}) ,$$

$$\kappa_4(T_n) = \frac{1}{n}\gamma_4 + O(n^{-2}) .$$

In the normal population case the distribution of r^2 and also the exact expressions of its moments are known, see e. g. Muirhead (1982). The leading terms $\gamma_0, \gamma_1, \ldots, \gamma_4$ of the expansions of the cumulants of r^2 in the normal population case are given by Lawley (1959). Hayakawa (1987) knows some of these expressions, namely $\gamma_0, \gamma_1, \gamma_3$ in the elliptical population case and has used them for the normalizing transformation of r^2 .

For evaluating $\gamma_0, \ldots, \gamma_4$ we use the expansion of r^2 through sample covariances, given by Lawley (1959):

$$r^2 = r_1 + r_2 + r_3 + \ldots ,$$

where

$$r_1 = 2\rho\bar{s}_{10} - \rho^2(\bar{s}_{00} + \bar{s}_{11}) ,$$

$$r_2 = \bar{s}_0^T\bar{s}_0 - 2\rho(\bar{s}_{00}\bar{s}_{10} + \bar{s}_0^T\bar{s}_1) + \rho^2(\bar{s}_{00}^2 + \bar{s}_{00}\bar{s}_{11} + \bar{s}_1^T\bar{s}_1) ,$$

$$r_3 = -\bar{s}_{00}\bar{s}_0^T\bar{s}_0 - \bar{s}_0^T\bar{S}_*\bar{s}_0 + 2\rho(\bar{s}_{00}^2\bar{s}_{10} + \bar{s}_{00}\bar{s}_0^T\bar{s}_1 + \bar{s}_0^T\bar{S}_*\bar{s}_1) -$$
$$- \rho^2(\bar{s}_{00}^3 + \bar{s}_{00}^2\bar{s}_{11} + \bar{s}_{00}\bar{s}_1^T\bar{s}_1 + \bar{s}_1^T\bar{S}_*\bar{s}_1) .$$

Here $\bar{s}_{ij} = s_{ij} - \sigma_{ij}$ and s_1 is the first column of S_* .

For calculating cumulants of r^2 the expressions of cumulants of sample covariances, given by Kaplan (1952), are needed. In the case of elliptical population Traat (1990) has got:

$$\kappa_2(s_{ij}, s_{kl}) = E(s_{ij} - \kappa_{ij})(s_{kl} - \kappa_{kl}) =$$

$$= \frac{1}{n}[k_4\kappa_{ij}\kappa_{kl} + (k_4 + 1)(\kappa_{ik}\kappa_{jl} + \kappa_{il}\kappa_{jk})] + O(n^{-2}) ,$$

$$\kappa_3(s_{ij}, s_{kl}, s_{ab}) = E(s_{ij} - \kappa_{ij})(s_{kl} - \kappa_{kl})(s_{ab} - \kappa_{ab}) =$$

$$= \frac{1}{n^2}[k_6\kappa_{ij}\kappa_{kl}\kappa_{ab} + (k_6 + 3k_4 + 1)\sum^{9}\kappa_{ik}\kappa_{ja}\kappa_{lb} +$$
$$+ (k_6 + 2k_4)\sum^{6}\kappa_{ij}\kappa_{ka}\kappa_{lb}] + O(n^{-3}) ,$$

where κ_{ij} is the element of population covariance matrix Σ. The fourth cumulant of sample covariances

$$\kappa_4(s_{ij}, s_{kl}, s_{ab}, s_{cd}) =$$
$$= E(s_{ij} - \kappa_{ij})(s_{kl} - \kappa_{kl})(s_{ab} - \kappa_{ab})(s_{cd} - \kappa_{cd}) - \kappa_2(s_{ij}, s_{kl})\kappa_2(s_{ab}, s_{cd}) -$$
$$\kappa_2(s_{ij}, s_{ab})\kappa_2(s_{kl}, s_{cd}) - \kappa_2(s_{ij}, s_{cd})\kappa_2(s_{kl}, s_{ab})$$

under an elliptical population is calculated with the help of the computer program made by Roomeldi (1991). It has 105 summands in its leading term.

We can simplify the derivation process carrying out the linear transformation of initial random vector x so, that its covariance matrix will take the form

$$\Sigma = cov(x) = \begin{bmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

The transformed $\mathbf{x}$ has still elliptical distribution and multiple correlation coefficient ρ. The distribution of r^2 does not change. The population covariances have the following values; $\varkappa_{10} = \varkappa_{01} = \rho$, $\varkappa_{ii} = 1$, $i = 0, \ldots, p$, other $\varkappa_{ij} = 0$.

Through straightforward but tedious calculations the following expressions for the leading terms of cumulants of statistic $T_n = \sqrt{n}(r^2 - \rho^2)$ are obtained:

$$\gamma_0 = 4c\rho^2(1 - \rho^2)^2 , \tag{9}$$

$$\gamma_1 = c(1 - \rho^2)(p - 2\rho^2) , \tag{10}$$

$$\gamma_2 = 2(1 - \rho^2)\{c^2(26\rho^4 - 2p\rho^2 + p) - 2(p + 5)\rho^2 - 2\rho^2[k_4(3p + 16) + k_6(p + 5)]\} , \tag{11}$$

$$\gamma_3 = 24c^2\rho^2(1 - \rho^2)^3(1 - 3\rho^2) , \tag{12}$$

$$\gamma_4 = 48\rho^2(1 - \rho^2)^4\{2c^3(1 - 2\rho^2)(2 - 13\rho^2) + \rho^2[3k_8 - 2k_4(6k_6 - 5k_4^2 - 6k_4)]\} , \tag{13}$$

where

$$c = k_4 + 1 .$$

From here follows the special case of normal population, where $c = 1$, $k_4 = k_6 = k_8 = 0$. Attention has to be paid to the term γ_2 which differs in a normal case from Lawley's result. We have to add the term $-4\rho^2(1 - \rho^2)$ to γ_2 to get the Lawley's result. Finally we have got the following result.

THEOREM. The analytical expression of the Edgeworth expansion of c.d.f. of statistic $T_n = \sqrt{n}(r^2 - \rho^2)$ under an elliptical population (1) is presented by expression (8) with $\gamma_0, \ldots, \gamma_4$ presented by (9) - (13).

3. Applications

In this section some figures are presented, where the simulated distribution of statistic T_n is compared with the different distributions: limiting normal distribution of T_n,

Edgeworth expansion up to the term $1/\sqrt{n}$, Edgeworth expansion up to the term $1/n$.

The 4-dimensional random vector with ε-contaminated normal distribution is observed:

$$x \sim EC_4(0, \Sigma, \psi) ,$$

where

$$\psi(t^T \Sigma t) = exp(-t^T \Sigma t/2) + (1 - \varepsilon) exp(-\sigma^2 t^T \Sigma t/2) .$$

The covariance matrix of x is

$$cov(x) = k_2 \Sigma ,$$

where

$$k_2 = \varepsilon + \sigma^2(1 - \varepsilon) .$$

Let Σ be of the form

$$\Sigma = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

then the square of multiple correlation coefficient of the vector x is ρ^2. The higher cumulants of x are presented by (2) - (4), where the expressions (5) - (7) of constants k_4, k_6, k_8 through ε and σ^2 are given by Traat (1990).

Let $\varepsilon = 0.8$ and $\sigma^2 = 9$, then

$$k_2 = 2.6, \quad k_4 = 1.515, \quad k_6 = 2.797, \quad k_8 = 0.574.$$

In the case $\varepsilon = 1$ we have the normal population, with $k_2 = 1$, $k_4 = k_6 = k_8 = 0$ and cov(x) = Σ.

Let sample size n = 9.

From the vector x 300 samples were generated, each one of size 9. From the sample covariance matrix the value of r^2 was calculated. The empirical c.d.f. of $T_n = \sqrt{n}(r^2 - \rho^2)$ was obtained.

On the figures the following notations are used:

———— simulated c.d.f. of T_n;

— — — limiting normal c.d.f. of T_n;

-------- first order Edgeworth expansion for T_n (i.e. up to the term $1/\sqrt{n}$);

..... second order Edgeworth expansion for the T_n (i.e. up to the term 1/n).

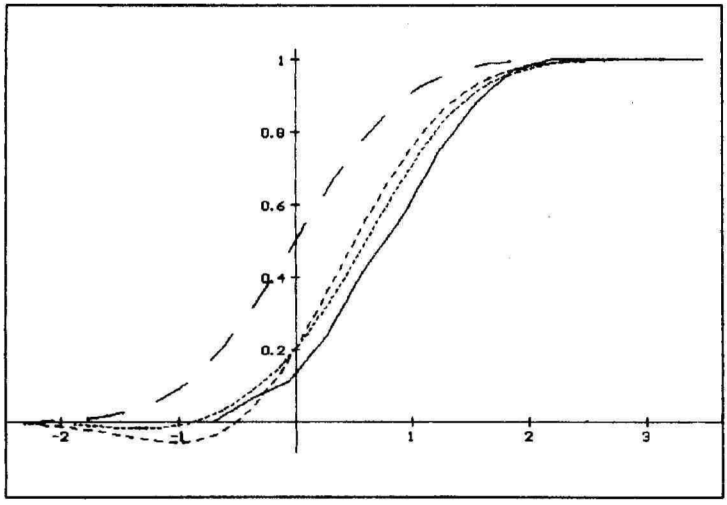

Figure 1. Normal population, $\rho = 0.5$, $n = 9$

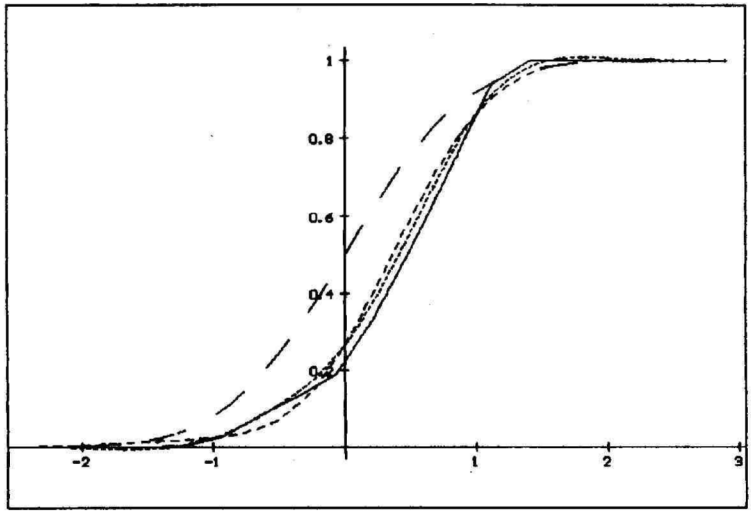

Figure 2. Normal population, $\rho = 0.7$, $n = 9$

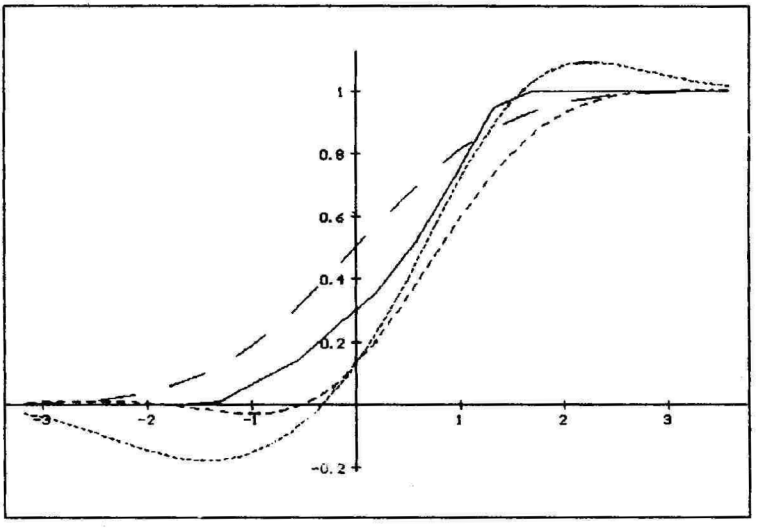

Figure 3. Normal contaminated population,
$\varepsilon = 0.8$, $\sigma^2 = 9$, $\rho = 0.7$, $n = 9$

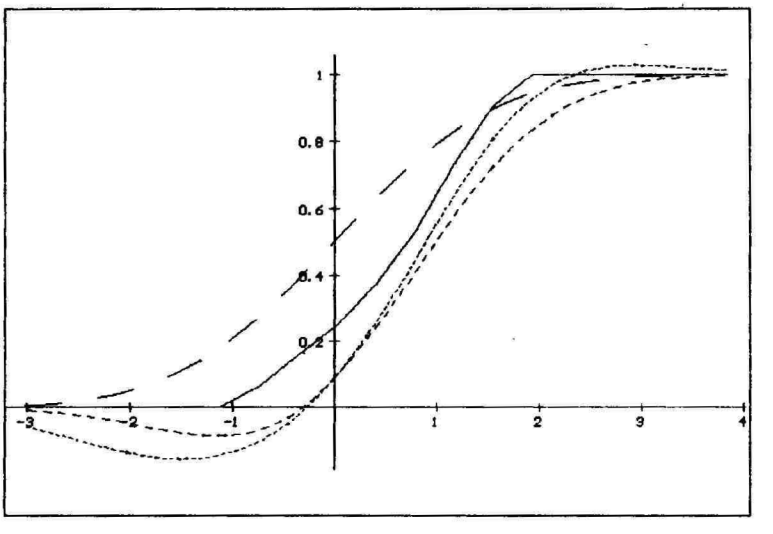

Figure 4. Normal contaminated population,
$\varepsilon = 0.8$, $\sigma^2 = 9$, $\rho = 0.6$, $n = 9$

In the normal population case the best approximation to the c.d.f. of T_n is second order Edgeworth expansion. In the case of elliptical population the approximation is not as good any more. But in the middle of the distribution the second order Edgeworth expansion still gives the best approximation to the c.d.f. of T_n.

On the figures 1) - 4) the medium values of ρ (0.5, 0.6, 0.7) were observed. If ρ is close to 0 or close to 1 then all approximating curves to c.d.f. of T_n behave very badly. The last phenomenon is presented on figure 5 for normal population.

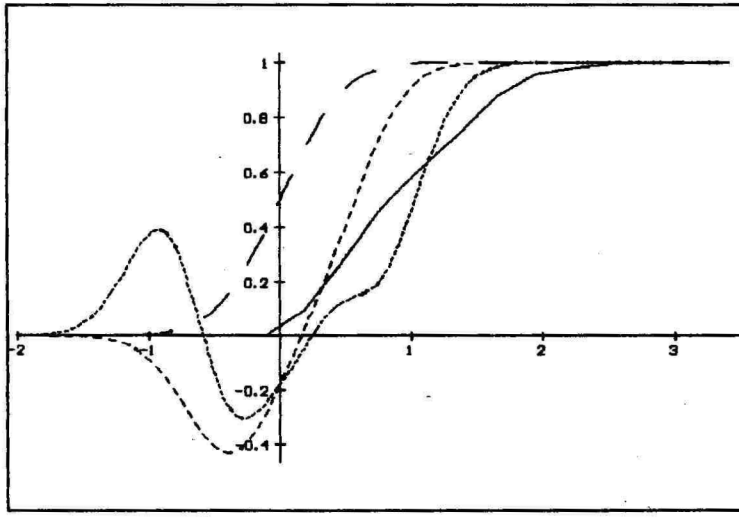

Figure 5. Normal population, ρ = 0.2, n = 9

References

Berkane, M. and Bentler, P. M. (1986). Moments of elliptically distributed random variates. Statist. and prob. Letters, 4, 333-335.

Bhattacharya, R. N., Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. Ann. Statist., 6, 434-451.

Fang, K. T., Kotz, S., NG, K. W. (1990). Symmetric multivariate and related distributions. London/New York: Chapman and Hall.

Hayakawa, T. (1987). Normalizing and variance stabilizing transformations of multivariate statistics under an elliptical population. Ann. Inst. Statist. Math., 39, 299-306.

Kaplan, E. L. (1952). Tensor notation and the sampling cumulants of k-statistics. Biometrika, 39, 319-323.

Lawley, D. N. (1959). Tests of significance in canonical analysis. Biometrika, 46, 59-66.

Muirhead, R. J. (1982). Aspects of multivariate statistical theory. New York: Wiley.

Roomeldi, R. (1991). Computer algebra in theory of elliptical population. Abstracts of International Conference on Algebra, Barnaul.

Traat, I. (1986). Matrix calculus for multivariate distributions. Acta Univ. Tartuensis, 733, 64-85.

Traat, I. (1990). Moments and cumulants of multivariate elliptical distribution with some applications. Acta Univ. Tartuensis, 893, 11-16.

Acta et Commentationes Universitatis Tartuensis, 942, 1992, p. 132–150.

ON ARRAYS AND ARRAY-DERIVATIVES *

Otto Karma
University of Tartu, Tartu, ESTONIA

Keywords: *Array, array-derivative, moments, characteristic function*

Introduction. In multivariate statistical analysis matrix differential calculus is used as a tool. Unfortunately there appear formal inconveniences connected with the arranging of the indices of multidimensional matrices, especially when derivatives of higher orders are considered. This artificially makes some formulas look more sophisticated than they are in essence and than they look in the one-dimensional case. For example the formula for derivative of Kronecker product $U \otimes V$ of matrices $U = U_{mxn}$, $V = V_{rxs}$ and $X = X_{pxq}$ is (see, e.g., Kollo (1991), paragraph 3.7, Magnus and Neudecker (1988), Section 9.14, Rodgers (1980), Theorem 6.6):
for vectorial derivative $\frac{dU}{dX} = \partial \mathrm{vec}U / \partial \mathrm{vec}'X$

$$\frac{d}{dX}(U \otimes V) = (I_n \otimes I_{m,s} \otimes I_r)[(I_{mn} \otimes \mathrm{vec}V)\frac{dU}{dX} + (\mathrm{vec}U \otimes I_{rs})\frac{dV}{dX}] \qquad (0.1)$$

and for derivative, arranged as block-matrix $\left[\frac{\partial u_{ij}}{\partial x_{kl}}\right]_{ik,jl}$,

$$\frac{\partial}{\partial X}(U \otimes V) = U\frac{\partial V}{\partial X} + (I_{r,m} \otimes I_p)(V \otimes \frac{\partial U}{\partial X})(I_{n,s} \otimes I_q). \qquad (0.2)$$

* The paper was submitted for publication in this volume by the Seminar of Tartu University on Statistics as a recent development of the results presented at the First Tartu Seminar on Multivariate Statistics.

Various approaches have been considered to avoid these formal inconveniences. We mention here only the idea of using the tensor calculus by Pollock (1985), which leads to results in some sense similar to ours.

We propose to use, instead of matrices, the sets of data, that depend on indices but do not depend on the arrangement of indices. We call these sets arrays. For the arrays $U = U_{MN}$, $V = V_{RS}$ and $X = X_{PQ}$ we get (see formula (5.6) in paragraph 5.3 below) the following analogue to the formulas (0.1), (0.2):

$$\frac{d}{dX}(U \otimes V) = U \otimes \frac{dV}{dX} + V \otimes \frac{dU}{dX} \quad . \tag{0.3}$$

Actually, sometimes the array is even a more natural way of presenting data than the matrix is. For example, let the following data characterize the distribution of the colour of hair $h \in H = \{d(dark), f(fair), o(other)\}$ and the nationality $n \in N = \{e(Estonian), o(other)\}$ in some group of people:

$$a_{d\in H,e\in N} = 12, \quad a_{f\in H,e\in N} = 23, \quad a_{o\in H,e\in N} = 10,$$
$$a_{d\in H,o\in N} = 20, \quad a_{f\in H,o\in N} = 9, \quad a_{o\in H,o\in N} = 15. \tag{0.4}$$

Then there are 24 natural ways to present these data as different matrices:

$$A_1 = \begin{pmatrix} 12 & 23 & 10 \\ 20 & 9 & 15 \end{pmatrix}, \quad A_2 = A_1^T, \quad A_3 = \begin{pmatrix} 12 & 10 & 23 \\ 20 & 15 & 9 \end{pmatrix}, \quad A_4 = A_3^T, \ldots .$$

At the same time these data are in a natural way presented by one array $A_{HN} = [a_{hn}]_{h\in H, n\in N}$ only. This array itself can be written down in the different forms, for example, as listing (0.4) or as follows:

A_{HN} =

H \ N	e	o
d	12	20
f	23	9
o	10	15

=

H \ N	d	f	o
e	12	23	10
o	20	9	15

=

H \ N	d	o	f
o	20	15	9
e	12	10	23

= ... .)

We also note that the multiplication of the arrays is commutative. That is why it may be convenient to use arrays even in the problems where only the algebraic calculations are needed.

At last we mention that the concept of array is in a very good accordance with the presentation of data in high-level computer languages. This may be important for practical calculations.

1. Arrays

1.1. In this paper we call *array* a set of elements depending on the indices $i_1^o \in I_1^o, \ldots, i_l^o \in I_l^o$ with $I_1^o, \ldots, I_l^o$ some index sets. At that for us an index set is the set of indices together with its notation (label), e.g., $I^o = \{I^o; 1,2,\ldots,9\}$. Therefore, the index sets $I^o = \{I^o; 1,2,\ldots,9\}$ and $J^o = \{J^o; 1,2,\ldots,9\}$ are different index sets although the sets of indices in these two index sets coincide. To show that the sets of indices of two index sets I^o and J^o coincide, we write $I^o (=) J^o$ and say that I^o and J^o are s-equal (equal as sets). In this case $i^o \in J^o$ for every index $i^o \in I^o$.

We assume that for every concrete array the index sets are fixed together with their notation and that

1) for every array the number l of its index sets is finite: $0 \le l < \infty$; the case $l = 0$ is considered as the array with the empty set of indices, the set of only one element without any indices,

2) for every index set I_r^o the number $|I_r^o|$ of the elements in I_r^o is finite: $1 \le |I_r^o| < \infty$,

3) all the indices of any array take their values from the different index sets; the only exception when two or more indices of the same array may take their values from the same index set is the case when the array is symmetric with respect to these indices - we explain it in a more detailed way in paragraph 1.4 below,

4) for every array, to any possible concrete set of values of indices corresponds exactly one element of the array.

Let us denote by $\emptyset$ the empty set and by N_n the index set $\{N_n; 1,2,\ldots,N_n\}$. In applications the index sets I_r^o are usually the sets $N_{n(r)}$, but it is not obligatory - they may consist of other elements and may also have their own structure. Some examples are: $I^o = \{I^o; -3,0,8,2\}$, $I^o = \{I^o; 1, \frac{1}{2}, \frac{1}{3}\}$, $I^o = \{I^o; (1;1),(1;2),(1;3),(2;1),(3;3)\}$.

1.2. We denote index sets by capital letters I, J,...

together with the obligatory superscript ^{o} and with optional indices, e.g., I^{o}, I^{o}_{1}, J^{o},... . As a rule, the corresponding symbols with small letters are used to denote the elements of the corresponding index sets and the values of these elements, e.g. $i^{o} \in I^{o}$, $i^{o}_{1} \in I^{o}_{1}$, $j^{o} \in J^{o}$,... .

The array of the elements, denoted by the letter a and with the indices $i^{o} \in I^{o},\dots,l^{o} \in L^{o}$, is denoted by $[a_{i^{o}\dots l^{o}}]_{i^{o}\in I^{o},\dots,l^{o}\in L^{o}}$ or $A_{I^{o}\dots L^{o}}$ or A. We do not list the membership of the indices ($i^{o} \in I^{o},\dots,l^{o} \in L^{o}$) if it is obvious from the notation of the indices or from the context, or if it is not important.

Two arrays with different index sets are always formally considered as different arrays. Thus $A_{I^{o}}$ and $A_{J^{o}}$ are two (formally) different arrays even if $I^{o}(=)J^{o}$.

For arrays it is formally not important in which order the indices and the membership of the indices (or the index sets) are listed. It may be important for convenience only. On the other hand, it *is* significant, from which index set every index takes its values. A concrete element of the array $A_{I^{o}J^{o}}$ is referred as $a_{i^{o}\in I^{o},j^{o}\in J^{o}}$, for example, as $a_{2\in I^{o},1\in J^{o}}$.

1.3. For convenience of notations we often denote all the list of the index sets of the array or only some part of this list by capital letters I,J,... without the superscript ^{o} but with optional indices, e.g., $I, I_{1},\dots$. As a rule, the corresponding small letters (with indices) are used to denote the corresponding multi-indices (lists of indices) or their values, e.g., if $I = I^{o}J^{o}$ then $i = i^{o}j^{o}$. Sometimes we also write a(i) instead of a_{i}.

We say that the lists I and J of index sets are equal and write I = J if I and J coincide or if they differ only in the ordering of index sets in these lists; for example, $I = J$ for $I = I^{o}_{1}I^{o}_{2}J^{o}$ and $J = I^{o}_{2}J^{o}I^{o}_{1}$. Obviously, if I = J and J = K then I = K.

If I = J, then we define the equality i = j of multi-indices $i \in I$ and $j \in J$ in the natural way as the equality of the indices from all the corresponding equal index sets.

We also use the following notations:

1) $I \subseteq J$, if $I = J$ or if the list I is equal to some part of the list J,

2) $K = I \cap J$ if K is the maximal list of index sets so that $K \subseteq I$, $K \subseteq J$,

3) $K = IJ$ if K is equivalent to the list of index sets that we get if we continue the list I with the list J,

4) $K = I \setminus J$ if $I = KL$ with $L = I \cap J$.

For the lists $I = I_1^o \dots I_l^o$ and $J = J_1^o \dots J_l^o$ we write $I (=) J$ and say that the lists I and J are s-equal if there exists one-to-one correspondence between the index sets of I and J so that corresponding index sets are s-equal. In this case for every multi-index $i \in I$ there exists corresponding multi-index $j \in J$ so that $i = j$; in this sense we can write $i \in J$ for every $i \in I$.

We note that using lists of index sets, which consist of only one index set, gives us the formal possibility to have different notations for one and the same index set. Sometimes it may be useful for explanations.

1.4. We say that the array A_I with $I = I_1^o \dots I_l^o$ and the array B_J are equal and write $A_I = B_J$ if $I = J$ and $a_{i_1^o \dots i_l^o} = b_{i_1^o \dots i_l^o}$ $\forall i_1^o \in I_1^o, \dots, i_l^o \in I_l^o$. Particularly, if $I = J$ then always $A_I = A_J$. Therefore, $A_I = B_J$ if and only if $A_I = B_I$. Obviously, if $A_I = B_I$ and $B_I = C_I$, then $A_I = C_I$.

We say that the array A_I with $I = I_1^o I_2^o J$ is symmetric with respect to indices from the index sets I_1^o and I_2^o (with respect to the indices $i_1^o \in I_1^o$, $i_2^o \in I_2^o$) if $I_1^o (=) I_2^o$ and the following *symmetry condition* is fulfilled:

$$\forall k, l \in I_1^o (=) I_2^o : a_{k \in I_1^o, l \in I_2^o, j \in J} = a_{k \in I_1^o, k \in I_2^o, j \in J} \qquad \forall j \in J.$$

The symmetry with respect to indices from more than two index sets and with respect to multi-indices are defined analogously.

If the array $A_{I_1 \dots I_l J}$ is symmetric with respect to the multi-indices $i_1 \in I_1, \dots, i_l \in I_l$, then it is often useful to work instead of this array with the array $A_{I_1 \dots I_l J}$, which differs from the array $A_{I_1 \dots I_l J}$ only formally. At that, if

the array A_I with $I = I_1 \ldots I_l J$ is considered, then always the following *consistency condition* for A_I must be fulfilled:

if we note for $A_{I_1 \ldots I_l J}$ *every list* I_s *differently, say* $I_1, I_2, \ldots, I_l$, *then the array* $A_{I_1 I_2 \ldots I_l J}$ *is symmetric with respect to multi-indices from the lists* $I_1, \ldots, I_l$.

(Otherwise condition 4) in paragraph 1.1 is not fulfilled.)

We note that, to avoid the difficulties connected with the consistency condition, it is always possible to use instead of the array A_{IJ} the array A_{IK} with $K(=)J$, $K \cap I = \varnothing$ and $a_{ik} = a_{ij}$ for $k = j$ $\forall i \in I$, $k \in K$, $j \in J$. (Actually we make thus the change of index sets, the arrays A_{IJ} and A_{IK} differ only formally.)

2. Array-product and Kronecker product of arrays

2.1. If it is not declared otherwise, we shall propose in this paper that the elements of the arrays are real or complex numbers or variables: $a_i \in \mathbb{K}$ with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Nevertheless, it is easy to see that they may be of much more general character.

For the constant $c \in \mathbb{K}$ and the arrays A_I and B_J with $I = J$ we define:

1) $cA_I = A_I c = R_I$ with $r_i = ca_i$ $\forall i \in I$, (2.1)

we also write (2.1) as follows: $c[a_i] = [a_i]c = [ca_i]$;

2) $A_I + B_J = A_I + B_I = R_I$ with $r_i = a_i + b_i$ $\forall_i \in I$, (2.2)

we also write (2.2) as follows: $[a_i] + [b_i] = [a_i + b_i]$.

The sum of more than two arrays is defined analogously.

It is easy to see that

1) $A = C$, $B = D \Rightarrow A + B = C + D$, $cA = cC$,

2) $A + B = B + A$, $(A + B) + C = A + (B + C) = A + B + C$,

3) $c'(c''A) = (c'c'')A$,

4) $(c' + c'')A = c'A + c''A$, $c(A + B) = cA + cB$.

2.2. In analogy with the matrices we define the array-product and the Kronecker product of the arrays.

For the two arrays A_I and B_J with $I \cap J = K$, $I_1 = I \setminus K$ $J_1 = J \setminus K$ we define the *array-product* $A_I B_J$ as the array $R_{I_1 J_1}$ with the elements

$$r_{i_1 j_1} = \begin{cases} \sum\limits_{k \in K} a_i b_j = \sum\limits_{k \in K} a_{i_1 k} b_{k j_1} & \text{if } K \neq \emptyset, \\ \sum\limits_{k \in \emptyset} a_{i_1 k} b_{k_1 j} = a_i b_j & \text{if } K = \emptyset. \end{cases} \tag{2.3}$$

As the special cases of the array-product we get analogues of $\circ$-product (scalar product) and $*$-product (star-product) of matrices:

$$A_I \circ B_I = A_I B_I, \quad A_I * B_{IJ} = A_I B_{IJ}. \tag{2.4}$$

For the two arrays A_I and B_J we define the *Kronecker product* $A_I \otimes B_J$ only if there exists the array R_{IJ} with the elements $r_{ij} = a_i b_j$; then

$$A_I \otimes B_J = R_{IJ}[a_i b_j]_{i \in I, j \in J}. \tag{2.5}$$

Thus, if $I = I_1 I_2$ and $J = I_2 J_1$, then the Kronecker product $A_I \otimes B_J$ exists only if

$$\forall k, l \in I_2: \; a_{i_1 k} b_{l j_1} = a_{i_1 l} b_{k j_1} \qquad \forall i_1 \in I_1, \; j_1 \in J_1,$$

i.e., if the consistency condition is fulfilled. We note that for existence of $A_I \otimes B_J$ it is sufficient that $I \cap J = \emptyset$.

The Kronecker product of more than two arrays we define analogously.

For every array A_I always exist the Kronecker products $A_I \otimes A_I$, $A_I \otimes A_I \otimes A_I$ etc. We denote $A_I \otimes \ldots \otimes A_I$ (l times A_I) by $A_I^{\otimes l}$ and agree that $A_I^{\otimes 0} = 1$. At that for $0 \leq r \leq l$:

$$A_I^{\otimes l} = [a_{i_1} \cdot \ldots \cdot a_{i_l}]_{i_1, \ldots, i_l \in I} =$$

$$= [(a_{i_1} \cdot \ldots \cdot a_{i_r})(a_{i_{r+1}} \cdot \ldots \cdot a_{i_l})] = A_I^{\otimes r} \otimes A_I^{\otimes l-r}.$$

2.3. We list some properties of the array-product and the Kronecker product of the arrays, which follow directly from the definitions:

1) $A_I = C_P$, $B_J = D_Q \;\rightarrow\; A_I B_J = C_P D_Q$,

$$\{\exists (A_I \otimes B_J) \;\rightarrow\; \exists (C_P \otimes D_Q), \quad A_I \otimes B_J = C_P \otimes D_Q\}$$

2) $A_I B_J = B_J A_I$;

$$\exists (A_I \otimes B_J) \;\rightarrow\; \exists (B_J \otimes A_I), \quad A_I \otimes B_J = B_J \otimes A_I, \tag{2.6}$$

3) $I = I_1 K_1$, $J = K_1 J_1 K_2$, $L = K_2 L_1$, $I \cap J = K_1$, $J \cap L = K_2 \;\rightarrow$ (2.7)

$$\to (A_I B_J)C_L = (A_{I K_1} B_{K_1 J K_2}) C_{K_2 L} = A_{I K_1}(B_{K_1 J K_2} C_{K_2 L}) = A_I (B_J C_L);$$

$$\exists((A_I \otimes B_J)\otimes C_L) \to \exists(A_I \otimes B_J \otimes C_L),\ (A_I \otimes B_J)\otimes C_L = A_I \otimes B_J \otimes C_L,$$

$$\exists(A_I \otimes (B_J \otimes C_L)) \to \exists(A_I \otimes B_J \otimes C_L),\ A_I \otimes (B_J \otimes C_L) = A_I \otimes B_J \otimes C_L, \qquad (2.8)$$

4) $c(A_I B_J) = (cA_I)B_J = A_I(cB_J)$;

$$\exists(A_I \otimes B_J) \to \exists(cA_I) \otimes B_J,\ \exists A_I \otimes (cB_J),$$

$$c(A_I \otimes B_J) = (cA_I) \otimes B_J = A_I \otimes (cB_J),$$

5) $A_I(B_J + C_J) = A_I B_J + A_I C_J = (B_J + C_J)A_I$;

$$\exists(A_I \otimes B_J),\ \exists(A_I \otimes C_J) \to \exists(A_I \otimes (B_J + C_J)),$$

$$A_I \otimes (B_J + C_J) = A_I \otimes B_J + A_I \otimes C_J,$$

6) $\exists(B_I \otimes C_J) \to (A_{IJK} B_I)C_J = A_{IJK}(B_I \otimes C_J)$, (2.9)

7) $I = I_1 K_1,\ J = J_1 K_2,\ P = K_1 P_1,\ Q = K_2 Q_1,\ I_1 J_1 \cap P_1 Q_1 = \varnothing,$
$\exists(A_I \otimes B_J),\ \exists(C_P \otimes D_Q),\ \exists((A_I C_P) \otimes (B_J D_Q)) \to$

$$\to (A_{I_1 K_1} \otimes B_{J_1 K_2})(C_{K_1 P_1} \otimes D_{K_2 Q_1}) = (A_{I_1 K_1} C_{K_1 P_1}) \otimes (B_{J_1 K_2} D_{K_2 Q_1}),$$

8) $\exists(A_I \otimes B_J) \to \exists(A_I \otimes B_J)^{\otimes l},\ (A_I \otimes B_J)^{\otimes l} = A_I^{\otimes l} \otimes B_J^{\otimes l}$, (2.10)

9) $I \cap J = \varnothing \to (A_{IK} B_{KJ})^{\otimes l} = A_{IK}^{\otimes l} B_{KJ}^{\otimes l}$, (2.11)

10) $$(A_I + B_I)^{\otimes l} = \sum_{r=0}^{l} \binom{l}{r} A_I^{\otimes r} \otimes B_I^{\otimes l-r}.$$ (2.12)

Let us prove here (2.11) and (2.12):

$$(A_{IK} B_{KJ}) \otimes \ldots \otimes (A_{IK} B_{KJ}) = [\sum_{k_1 \in K} a_{i k_1} b_{k_1 j}] \otimes \ldots \otimes [\sum_{k_l \in K} a_{i k_l} b_{k_l j}] =$$

$$= [\sum_{k_1 \in K} a_{i_1 k_1} b_{k_1 j_1} \cdot \ldots \cdot \sum_{k_l \in K} a_{i_l k_l} b_{k_l j_l}]_{i_1, \ldots, i_l \in I,\ j_1, \ldots, j_l \in J} =$$

$$= [\sum_{k_1, \ldots, k_l \in K} (a_{i_1 k_1} \ldots a_{i_l k_l})(b_{k_1 j_1} \ldots b_{k_l j_l})] = A_{IK}^{\otimes l} B_{KJ}^{\otimes l};$$

$$(A_I + B_I) \otimes \ldots \otimes (A_I + B_I) = [(a_{i_1} + b_{i_1}) \cdot \ldots \cdot (a_{i_l} + b_{i_l})]_{i_1, \ldots, i_l \in I} =$$

$$= [a_{i_1} \cdot \ldots \cdot a_{i_l}] + [a_{i_1} \cdot \ldots \cdot a_{i_{l-1}} b_{i_l}] +$$

$$+ [a_{i_1} \cdot \ldots \cdot a_{i_{l-2}} b_{i_{l-1}} a_{i_l}] + \ldots + [b_{i_1} a_{i_2} \cdot \ldots \cdot a_{i_l}] +$$

$$+ [a_{i_1} \cdot \ldots \cdot a_{i_{l-2}} b_{i_{l-1}} b_{i_l}] + [a_{i_1} \cdot \ldots \cdot a_{i_{l-3}} b_{i_{l-2}} a_{i_{l-1}} b_{i_l}] + \ldots$$

$$\ldots + [b_{i_1} b_{i_2} a_{i_3} \cdot \ldots \cdot a_{i_l}] + \ldots + [b_{i_1} \cdot \ldots \cdot b_{i_l}] =$$

$$= A_I^{\otimes l} + 1 A_I^{\otimes l-1} \otimes B_I + \binom{l}{2} A_I^{\otimes l-2} \otimes B_I^{\otimes 2} + \ldots + B_I^{\otimes l} = \sum_{r=0}^{l} \binom{l}{r} A_I^{\otimes r} \otimes B_I^{\otimes l-r}.$$

3. Linear vector space of arrays

3.1. All the arrays with the same list I of the index sets form a finite-dimensional linear vector space over $\mathbb{K}$. We denote this space by $\mathcal{U}_I$. If $I = I_1^0 \dots I_l^0$, then $\dim \mathcal{U}_I \le |I_1^0| \cdot \dots \cdot |I_l^0|$, particularly if all the index sets $I_1^0, \dots, I_l^0$ are pairwise different, then $\dim \mathcal{U}_I = |I_1^0| \cdot \dots \cdot |I_l^0|$.

The vector space $\mathcal{U}_I$ can be considered as a Hilbert space with the scalar product $(A_I, B_I) = A_I \overline{B}_I$ where $\overline{B}_I = [\overline{b}_i]$ with $\overline{b}_i$ the conjugate to b_i. It also can be also considered as a Banach space with respect to various norms, for example,

$$\|A_I\|_0 = \max_{i \in I} |a_i|, \quad \|A_I\|_2 = \Big(\sum_{i \in I} |a_i|^2 \Big)^{1/2} = (A_I, \overline{A}_I)^{1/2}.$$

At that all the norms in A_I are equivalent whereas $\dim \mathcal{U}_I < \infty$. Only for concreteness we fix one norm in $\mathcal{U}_I$, say $\|\cdot\|_0$. Then such notions as neighbourhood, etc. have a concrete meaning.

3.2. The arrays $L_{IJ} \in \mathcal{U}_{IJ}$ can be considered as linear operators $L \in \mathcal{B}(\mathcal{U}_I, \mathcal{U}_J)$ from $\mathcal{U}_I$ to $\mathcal{U}_J$ defined by the equality $L_{IJ}(A_I) = L_{IJ} A_I$. (Obviously, as well $L_{IJ} \in \mathcal{B}(\mathcal{U}_J \mathcal{U}_I)$ with $L_{IJ}(A_J) = L_{IJ} A_J$.) Besides, it is not difficult to see that if $I \cap J = \varnothing$ then all the linear operators $L \in \mathcal{B}(\mathcal{U}_I, \mathcal{U}_J)$ can be represented by the arrays $L_{IJ} \in \mathcal{U}_{IJ}$ whereas $\dim \mathcal{U}_{IJ} = \dim \mathcal{U}_I \cdot \dim \mathcal{U}_J = \dim \mathcal{B}(\mathcal{U}_I, \mathcal{U}_J)$.

In the special case of $J = \varnothing$ we get $\mathcal{B}(\mathcal{U}_I, \mathbb{K}) = \mathcal{U}_I^*$, the conjugate space to $\mathcal{U}_I$. Thus every linear functional on $\mathcal{U}_I$ can be represented by array-multiplication with the certain element of $\mathcal{U}_I$ itself.

We also note that every array $L_I \in \mathcal{U}_I$ can be considered as the linear operator $L \in \mathcal{B}(\mathcal{U}_{IJ}, \mathcal{U}_J)$ defined by $L_I(A_{IJ}) = L_I A_{IJ}$, or as the linear operator $L \in \mathcal{B}(\mathcal{U}_I, \mathcal{U}_{IJ})$ defined by $L_I(A_{IJ}) = L_I \circledast A_J$, if the Kronecker product $L_I \circledast A_J$ exists.

4. Array-derivative

4.1. Let us consider the array U_P, which is the function of the array $X_I \in \mathcal{U}_I$ in some region (open connected set) $D \subseteq \mathcal{U}_I$, i.e., all the elements u_p, $p \in P$ of the array U_P are the functions of all the elements x_i, $i \in I$ for $[x_i]_{i \in I} = X_I \in D$. At that x_i, $i \in I$ are considered as independent variables.

If all the partial derivatives $\dfrac{\partial u_p(X_I)}{\partial x_i} = \dfrac{\partial u_p}{\partial x_i} X_I$, $i \in I$,

$p \in P$, $X_I \in D$ and the array $\left[\frac{\partial u_p}{\partial x_i}\right]_{pI}$ exist, then we denote the array $\left[\frac{\partial u_p}{\partial x_i}\right]_{pI}$ by $\frac{dU_p}{dX_I}$ and call it the *array-derivative* of U_p by X_I in D or simply the derivative of U_p by X_I. (We note that if all the partial derivatives $\frac{\partial u_p}{\partial x_i}$ exist, then for existence of the array $\frac{dU_p}{dX_I}$ it is sufficient that $P \cap I = \emptyset$.)

We say that U_p is differentiable by X_I at $X_I^* \in D$ if there exists $\frac{dU_p(X_I^*)}{dX_I}$ and if all the functions u_p, $p \in P$ are differentiable at X_I^*:

$$u_p(X_I^* + \Delta X_I) = u_p(X_I^*) + \sum_{i \in I} \frac{\partial u_p}{\partial x_i}(X_I^*)\Delta x_i + \alpha_p(X_I^*, \Delta X_I), \quad \forall p \in P, \qquad (4.1)$$

where $\Delta X_I = [\Delta x_i]_{i \in I}$ and $|\alpha_p(X_I^*, \Delta X_I)|/\|\Delta X_I\| \to 0$ if $\|\Delta X_I\| \to 0$.

We recall that for (4.1) it is sufficient that all the partial derivatives $\frac{\partial u_p}{\partial x_i}$ are continuous functions of X_I at X_I^*.

If U_p is differentiable by X_I at X_I^*, then

$$\Delta U_p(X_I^*) = U_p(X_I^* + \Delta X_I) - U_p(X_I^*) = [u_p(X_I^* + \Delta X_I) - u_p(X_I^*)] =$$

$$= [\sum_{i \in I} \frac{\partial u_p}{\partial x_i}(X_I^*)\Delta x_i + \alpha_p(X_I^*, \Delta X_I)] = \frac{dU_p(X_I^*)}{dX_I}\Delta X_I + \alpha_p(X_I^*, \Delta X_I) \qquad (4.2)$$

with $\alpha_p(X_I^*, \Delta X_I) = [\alpha_p(X_I^*, \Delta X_I)]$, $\|\alpha_p(X_I^*, \Delta X_I)\|/\|\Delta X_I\| \to 0$ if $\|\Delta X_I\| \to 0$.

4.2. Let all $x_i, i \in I$ be independent variables. From the corresponding properties of the derivatives of functions we immediately get for U and V depending on X and for A not depending on X:

1) $\frac{dA_p}{dX_I} = \left[\frac{\partial a_p}{\partial x_i}\right] = O_{pI}$ with O_{pI} the array of zeroes,

2) for I (=) J:

$$\frac{dX_I}{dX_J} = \left[\frac{\partial x_i}{\partial x_j}\right] = E_{IJ} \text{ with } E_{IJ} = [\delta_{ij}]_{i \in I, j \in J},$$

where δ_{ij} is the Kronecker symbol, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

3) $\exists \frac{dU_P}{dX_I} \Rightarrow \exists \frac{d(cU_P)}{dX_I}, \quad \frac{d(cU_P)}{dX_I} = c\frac{dU_P}{dX_I}, \quad \forall c \in \mathbb{K},$

4) $\exists \frac{dU_P}{dX_I}, \frac{dV_P}{dX_I} \Rightarrow \exists \frac{d(U_P+V_P)}{dX_I}, \quad \frac{d(U_P+V_P)}{dX_I} = \frac{dU_P}{dX_I} + \frac{dV_P}{dX_I}$

5) $\exists \frac{dU_{QP}}{dX_I}, \quad R \cap I = \emptyset \Rightarrow$

$$\Rightarrow \exists \frac{d(A_{RQ}U_{QP})}{dX_I}, \quad \frac{d(A_{RQ}U_{QP})}{dX_I} = A_{RQ}\frac{dU_{QP}}{dX_I}. \qquad (4.3)$$

At that if U and V are differentiable by X at X_I^*, then cU, U + V and AU are also differentiable by X at X_I^*.

Although the following properties are the special cases of (4.3), it is easier to verify them directly:

6) for I (=) J, P $\cap$ J = $\emptyset$:

$$\frac{d(A_{PI}X_I)}{dX_J} = \left[\frac{d}{dx_j}\sum_{i\in I} a_{pi}x_i\right] = [a_{pj}] = A_{PJ}, \qquad (4.4)$$

7) for I (=) J, Q = S, R $\cap$ I = $\emptyset$:

$$\frac{d(A_{RQ}X_{QI})}{dX_{SJ}} = \left[\frac{d}{dx_{sj}}\sum_{q\in Q} a_{rq}x_{qi}\right] = [a_{rs}\delta_{ij}] = A_{RS} \otimes E_{IJ}, \qquad (4.5)$$

8) for I = J, R $\cap$ I = $\emptyset$:

$$\frac{d(A_R \otimes X_I)}{dX_J} = \left[\frac{d}{dx_j} a_r x_i\right] = [a_r\delta_{ij}] = A_R \otimes E_{IJ}. \qquad (4.6)$$

4.3. Finding the array-derivative of $U_P \in \mathfrak{U}_P$ by X_I at X_I^*, if it exists, can be considered as an application of a linear operator $L_I \in \mathfrak{B}(\mathfrak{U}_P, \mathfrak{U}_{PI})$ to U_P. We denote this operator by $\frac{d}{dX_I}\Big|_{X_I^*}$. Then $\frac{d}{dX_I}\Big|_{X_I^*} U_P = \frac{dU_P^*(X_I)}{dX_I}$.

If U_P is differentiable by X_I at X_I^*, then by (4.2)

$$U_P(X_I^* + \Delta X_I) = U_P(X_I^*) + \left(\frac{d}{dX_I}\Big|_{X_I^*} U_P\right)\Delta X_I + \alpha_P(X_I^*, \Delta X_I) \qquad (4.7)$$

with $\|\alpha_P(X_I^*, \Delta X_I)\|/\|\Delta X_I\| \to 0$ if $\|\Delta X_I\| \to 0$, i.e., U is Frechet' differentiable at X_I^* and $\frac{d}{dX_I}\Big|_{X_I^*} U_P = \frac{dU_P(X_I^*)}{dX_I} \in \mathfrak{B}(\mathfrak{U}_I, \mathfrak{U}_P)$ is the Frechet' derivative $U'(X^*)$ of the function U_P at X_I^*.

Vice versa, if U_P is Frechet' differentiable at X_I^*, i.e., if there exists $U'(X^*) \in \mathfrak{B}(\mathfrak{U}_I, \mathfrak{U}_P)$ so that

$$U_p(X_I^* + \Delta X_I) = U_p(X_I^*) + U'(X^*)\Delta X_I + \alpha_p(X_I^*, \Delta X_I) \quad \forall \Delta X_I \in \mathcal{U}_I$$

with $\|\alpha_p(X_I^*, \Delta X_I)\|/\|\Delta X_I\| \to 0$ if $\|\Delta X_I\| \to 0$, then

$$u_p(X_I^* + \Delta X_I) = u_p(X_I^*) + \sum_{i \in I} l_{pi}\Delta X_i + \alpha_p(X_I^*, \Delta X_I), \quad \forall \Delta X_I \in \mathcal{U}_I, \quad \forall p \in P,$$

from which it follows that every function u_p is differentiable at X_I^* and $l_{pi} = \frac{\partial u_p}{\partial x_i}(X_I^*)$. Thus, if $\frac{dU_p(X_I^*)}{dX_I}$ exists then $U'(X^*)\Delta X = \frac{dU_p(X_I^*)}{dX_I}\Delta X_I$ and the Frechet' derivative of U_p at X_I^* is equal to $\frac{dU_p(X_I^*)}{dX_I} \in \mathcal{B}(\mathcal{U}_I, \mathcal{U}_p)$.

4.4. Let the array $U_{P_{|I|}}$ be the function of the array X_I which in turn is the function of the array T_L, so that U_p is actually the function of T_L. If there exists $\frac{dx_i}{dt_l}$ $\forall i \in I$ at T_L^* and if all u_p, $p \in P$ are differentiable by X_I at $X_I^* = X_I(T_L^*)$, then there exist all the partial derivatives of u_p by t_l and

$$\frac{\partial u_p}{\partial t_l}(T_L^*) = \sum_{i \in I} \frac{\partial u_p}{\partial x_i}(X_I^*)\frac{\partial x_i}{\partial t_l}(T_L^*).$$

Therefore, if there exist arrays $\frac{dX_I}{dT_L}$ and $\frac{dU_p}{dT_L}$ at T_L^* and if U_p is differentiable by X_I at $X_I^* = X_I(T_L^*)$, then

$$\frac{dU_p}{dT_L}(T_L^*) = \left[\sum_{i \in I} \frac{\partial u_p}{\partial x_i}(X_I^*)\frac{\partial x_i}{\partial t_l}(T_L^*)\right] = \frac{dU_p(X_I^*)}{dX_I}\frac{dX_I(T_L^*)}{dT_L}. \qquad (4.8)$$

At that, if X_I is differentiable at T_L^*, then U_p is differentiable at T_L^*.

As the application of (4.8), let us find the derivative of the function $e^{T_I A_I}$ by T_I, considering A_I as parameters-constants. Let us denote $y(T_I) := T_I A_I$. Then, for $I = J$, with $y_o = y(T_I^*) = T_I^* A_I$,

$$\frac{de^{T_I A_I}(T_I^*)}{dT_J} \overset{(4.8)}{=} \frac{de^{y}(y_o)}{dy}\frac{dy(T_I^*)}{dT_J} \overset{(4.4)}{=} e^{y_o}A_J = A_J e^{T_I^* A_I}$$

and therefore

$$\frac{de^{T_I A_I}}{dT_I} = A_I e^{T_I A'_I}. \qquad (4.9)$$

5. Derivatives of higher orders and partial derivatives

5.1. If there exists the array $U' = \frac{dU_P}{dX_I}$, then it is a function of X_I. If all the partial derivatives $\frac{\partial u_p}{\partial x_i}$ are differentiable, then $\frac{\partial^2 u}{\partial x_{i_1} \partial x_{i_2}} = \frac{\partial^2 u}{\partial x_{i_2} \partial x_{i_1}}$ and U' is differentiable by X. We call the derivative of U' by X_I the second-order array-derivative of U_P by X_I and denote it by $\frac{d^2 U_P}{dX_I^2}$. Thus

$$\frac{d^2 U_P}{dX_I^2} := \frac{d}{dX_I}\left(\frac{dU_P}{dX_I}\right) = \left[\frac{\partial}{\partial x_{i_1}} \frac{\partial u_p}{\partial x_{i_2}}\right] = \left[\frac{\partial^2 u_p}{\partial x_{i_1} \partial x_{i_2}}\right]_{p \in P, i_1, i_2 \in I}.$$

At that $\frac{d^2 U_P}{dX_I^2} \in \mathfrak{B}(\mathfrak{U}_I, \mathfrak{B}(\mathfrak{U}_I, \mathfrak{U}_P))$ is the second-order Frechet' derivative U'' of U_P by X_I and

$$(U''\Delta X)\Delta X = \left(\frac{d^2 U_P}{dX_I^2}\Delta X_I\right)\Delta X_I \quad (2.5) \quad \frac{d^2 U_P}{dX_I^2} \Delta X_I^{\otimes 2}$$

The derivatives of U_P by X_I of higher orders are defined analogously:

if $\frac{d^{l-1} U_P}{dX_I^{l-1}}$ is differentiable by X_I, then

$$\frac{d^l U_P}{dX_I^l} := \frac{d}{dX_I}\left(\frac{d^{l-1} U_P}{dX_I^{l-1}}\right) = \left[\frac{\partial^l u_p}{\partial x_{i_1} \ldots \partial x_{i_l}}\right]_{p \in P, i_1, \ldots, i_l \in I}; \qquad (5.1)$$

at that $U^{(l)} = \frac{d^l U_P}{dX_I^l} \in \mathfrak{B}(\mathfrak{U}_I, \mathfrak{B}(X, \ldots, \mathfrak{B}(\mathfrak{U}_I, \mathfrak{U}_P)\ldots))$ (l times $\mathfrak{U}_I$) is the l-th Frechet' derivative $U^{(l)}$ of U_P by X_I and

$$(\ldots((U^{(l)}\Delta X)\Delta X\ldots)\Delta X = (\ldots(\left(\frac{d^l U_P}{dX_I^l}\right)\Delta X_I)\Delta X_I \ldots \Delta X_I)\Delta X_I =$$

$$= \frac{d^l U_P}{dX_I^l} \Delta X_I^{\otimes l}. \qquad (5.2)$$

Thus we can use the general results of functional ana-

lysis for Banach spaces. For example, we get the formula of Taylor (with the Peano reminder term):

if U_P is l times differentiable by X_I at X_I^*, then

$$U_P(X_I^* + \Delta X_I) = U_P(X_I^*) + \frac{dU_P(X_I^*)}{dX_I}\Delta X_I + \frac{d^2U_P(X_I^*)}{dX_I^2}\Delta X_I^{\otimes r} + \\ + \ldots + \frac{d^lU_P(X_I^*)}{dX_I^l}\Delta X_I^{\otimes l} + \alpha_{Pl}(X_I^*, \Delta X_I) \qquad (5.3)$$

with $\|\alpha_{Pl}(X_I^*, \Delta X_I)\|/\|\Delta X_I\| \to 0$ if $\|\Delta X_I\| \to 0$.

5.2. Let the array U_P be a function of two arrays X_I and Y_J. Then, instead of $\frac{dU_P}{dX_I}$, we write $\frac{\partial U_P}{\partial X_I}$ and call it the partial array-derivative of U_P by X_I. We say that U_P is differentiable at (X_I^*, Y_J^*) if all the functions $u_p(X_I, Y_J)$ are differentiable at (X_I^*, Y_J^*) and if there exist the arrays $\frac{\partial U_P(X_I^*, X_J^*)}{\partial X_I}$ and $\frac{\partial U_P(X_I^*, X_J^*)}{\partial Y_J}$; it is not difficult to see that then

$$U_P(X_I^* + \Delta X_I, Y_J^* + \Delta Y_J) = U_P(X_I^*, Y_J^*) + \frac{\partial U_P(X_I^*, X_J^*)}{\partial X_I}\Delta X_I + \\ + \frac{\partial U_P(X_I^*, X_J^*)}{\partial Y_J}\Delta Y_J + \alpha_P(X_I^*, Y_J^*, \Delta X_I, \Delta Y_J) \qquad (5.4)$$

with $\|\alpha_P(X_I^*, Y_J^*, \Delta X_I, \Delta Y_J)\|/(\|\Delta X_I\| + \|\Delta Y_J\|) \to 0$ if $\|\Delta X_I\| + \|\Delta Y_J\| \to 0$.

The mixed partial derivatives of higher orders can be introduced analogously in a natural way and the case of more than two array-arguments also can be considered but we do not consider these problems in this paper.

5.3. Let U_P be the function of two arrays X_I and Y_J, which in turn are the functions of the arrays S_K and T_L. If there exist partial derivatives $\frac{\partial X_I(S_K^*, T_L^*)}{\partial S_K}$ and $\frac{\partial Y_J(S_K^*, T_L^*)}{\partial S_K}$ and if U_P is differentiable at (X_I^*, Y_J^*) with $X_I^* = X_I(S_K^*, T_K^*)$, $Y_J^* = Y_J(S_K^*, T_K^*)$, then there exist partial derivatives

$$\frac{\partial u_p(S_K^*, T_L^*)}{\partial s_k} = \sum_{i \in I} \frac{\partial u_p(X_I^*, Y_J^*)}{\partial x_i}\frac{\partial x_i(S_K^*, T_L^*)}{\partial s_k} +$$

$$+ \sum_{j \in J} \frac{\partial u_p(X_I^*, Y_J^*)}{\partial y_j} \frac{\partial y_j(S_K^*, T_L^*)}{\partial s_k}.$$

If at that there exists the array $\frac{\partial U_P(S_K^*, T_L^*)}{\partial S_K}$, then

$$\frac{\partial U_P(S_K^*, T_L^*)}{\partial S_K} = \frac{\partial U_P(X_I^*, Y_J^*)}{\partial X_I} \frac{\partial X_I(S_K^*, T_L^*)}{\partial S_K} + \frac{\partial U_P(X_I^*, Y_J^*)}{\partial Y_J} \frac{\partial Y_J(S_K^*, T_L^*)}{\partial S_K}. \tag{5.4}$$

As the application of (5.4) we find the derivative of UV and $U \circledast V$ if U and V are the functions of X:

a) Let $U = U_{RK}$, $V = V_{KS}$ and $X = X_I$ with $R \cap I = S \cap I = K \cap I = R \cap S = R \cap K = K \cap S = \emptyset$. Then

$$\frac{d(UV)}{dX} = \frac{d(U_{RK}V_{KS})}{dX_I} \overset{(5.4)}{=} \frac{\partial(V_{SK}U_{KR})}{\partial U_{KR}} \frac{dU_{RK}}{dX_I} + \frac{\partial(U_{RK}V_{KS})}{\partial V_{KS}} \frac{dV_{KS}}{dX_I} \overset{(4.5)}{=}$$

$$= (V_{KS} \circledast E_{RR}) \frac{dU_{RK}}{dX_I} + (U_{RK} \circledast E_{SS}) \frac{dV_{KS}}{dX_I} = \tag{5.5}$$

$$= V_{KS} \frac{dU_{RK}}{dX_I} + U_{RK} \frac{dV_{KS}}{dX_I} = V \frac{dU}{dX} + U \frac{dV}{dX}.$$

b) Let $U = U_R$, $V = V_S$ and $X = X_I$ with $R \cap S = R \cap I = S \cap I = \emptyset$. Then

$$\frac{d(U \circledast V)}{dX} = \frac{d(U_R \circledast V_S)}{dX_I} \overset{(5.4)}{=} \frac{\partial(U \circledast V)}{\partial U} \frac{dU}{dX} + \frac{\partial(U \circledast V)}{\partial V} \frac{dV}{dX} \overset{(4.6)}{=}$$

$$= (V_S \circledast E_{RR}) \frac{dU_R}{dX_I} + (U_R \circledast E_{SS}) \frac{dV_S}{dX_I} = \tag{5.6}$$

$$= V_S \frac{dU_R}{dX_I} + U_R \frac{dV_S}{dX_I} = V \circledast \frac{dU}{dX} + U \circledast \frac{dV}{dX}.$$

6. Random arrays

6.1. Analogously to the cases of vectors and matrices, we call *random array* the array, which elements are random variables. We consider in this paper only some connections between the moments and the characteristic function of the real random array to see that they have formally the same form as for random variables.

We denote the mean of the random variable x_i by Ex_i. If for the random array X_I for every x_i, $i \in I$ there exists (i.e., is finite) the mean Ex_i then we call the array

$[Ex_i]_{i\in I}$ the mean of the array X_I and denot it by EX_I.

6.2. If for the random array X_I and the array A_I for every set of indices $\{i_1,\dots,i_l\}$ with $i_1,\dots,i_l \in I$, $l \geq 1$ there exists the mean $E((x_{i_1} - a_{i_1})\cdot\dots\cdot(x_{i_l} - a_{i_l}))$, then the array $[E((x_{i_1} - a_{i_1})\cdot\dots\cdot(x_{i_l} - a_{i_l}))]_{i_1,\dots,i_l\in I}$ is called the array of the l-th moments of X_I with respect to A_I and denoted by $m_A^{(l)}(X)$ or m_A^l. (We recall that if there exist all the means $E|x_i|^l$, $i \in I$, then there also exist $E|x_i|^r$, $E(|x_{i_1} - a_{i_1}|\cdot\dots\cdot|x_{i_r} - a_{i_r}|)$ and $E((x_{i_1} - a_{i_1})\cdot\dots\cdot(x_{i_r} - a_{i_r}))$ for $i, i_1,\dots,i_r \in I$, $1 \leq r \leq l$.) As usual, we agree that $m_A^{(0)}(X) = 1$.

The most important particular cases of $m_A^{(l)}(X)$ are $A_I = 0_I = [0_i]$ and $A_I = EX_I$; the array $m_0^{(l)}(X)$ is denoted by $m^{(l)}(X)$ and called the array of the l-th moments of X, the array $m_{EX}^{(l)}(X)$ is denoted by $\mu^{(l)}(X)$ and called the array of the l-th central moments of X.

We note that

$$m_A^{(l)}(X) = [E((x_{i_1} - a_{i_1})\cdot\dots\cdot(x_{i_l} - a_{i_l}))]_{i_1,\dots,i_l\in I} \quad (6.1)$$

$$= E((X_I - A_I)\otimes\dots\otimes(X_I - A_I)) = E(X - A)^{\otimes l} = m_0^{(l)}(X - A),$$

i.e., the array $m_A^{(l)}(X)$ is equal to the array of the l-th moments of the array $X - A$ and to the mean of the array $(X - A)^{\otimes l}$.

6.3. From the properties of stochastic variables it follows easily that for stochastic arrays X_I and Y_I:

1) $EB_I = [Eb_i] = B_I$ for every B_I with $b_i \in \mathbb{R}$,

2) $\exists EX_I \Rightarrow \exists E(cX_I)\ \forall c \in \mathbb{R}$, $E(cX_I) = [E(cx_j)] = cEX_I$,

3) $\exists EX_I, \exists EY_I \Rightarrow \exists E(X_I + Y_I)$, $E(X_I + Y_I) = EX_I + EY_I$,

4) $\exists EX_I$, $b_{ki} \in \mathbb{R} \Rightarrow \exists E(B_{KI}X_I)$,

$$E(B_{KI}X_I) = [E\sum_{i\in I} b_{ki}x_i] = [\sum_{i\in I} b_{ki}Ex_i] = B_{KI}EX_I,$$

5) $\exists EX_I$, $b_k \in \mathbb{R} \Rightarrow \exists E(B_K \otimes X_I)$,

$$E(B_K \otimes X_I) = [E(b_k x_i)] = [b_k Ex_i] = B_K \otimes EX_I.$$

Therefore from (6.1) it follows that if there exist $E|x_i|^l$, $E|y_i|^l$, $i \in I$, then:

1) $m_A^{(l)}(B) = E(B - A)^{\otimes l} = (B - A)^{\otimes l}$,

2) $\exists m^{(l)}_{cA}(cX),\ m^{(l)}_{cA}(cX) = E(cX-cA)^{\otimes l} = c^l E(X-A)^{\otimes l} = c^l m^{(l)}_A(X),$

3) $\exists m^{(l)}_{A+B}(X+Y),\ m^l_{A+B}(x+y) = E(X-A+Y-B)^l \underset{(2.12)}{=}$

$$= \sum_{r=0}^{l} \binom{l}{r} E((X-A)^{\otimes r} \otimes (Y-B)^{\otimes l-r}),$$

4) $\exists m^{(l)}_{BA}(BX),\ m^{(l)}_{BA}(BX) = E(B_{KI}X_I - B_{KI}A_I)^{\otimes l} = E(B_{KI}(X_I-A_I))^{\otimes l} \underset{(2.14)}{=}$

$$= E(B^{\otimes l}_{KI}(X_I - A_I)^{\otimes l}) = B^{\otimes l}_{KI} E(X-A)^{\otimes l} = B^{\otimes l}_{IK}\, m^{(l)}_A(X)$$

5) $\exists (B_J \otimes X_I) \to \exists m^{(l)}_{B\otimes A}(B\otimes X),\ m^{(l)}_{B\otimes A}(B\otimes X) = E(B\otimes X - B\otimes A)^{\otimes l} =$

$$= E(B\otimes(X-A))^{\otimes l} \underset{(2.10)}{=} E(B^{\otimes l}\otimes(X-A)^{\otimes l}) = B^{\otimes l}\otimes E(X-A)^{\otimes l} =$$

$$= B^{\otimes l}\otimes m^{(l)}_A(X)$$

(particularly $m^{(2)}(B_JX_I) = B^{\otimes 2}_J m^{(2)}(X_I) = B_J \otimes m^{(2)}(X_I) \otimes B_J$).

Here we also note that

$$m^{(l)}_A(X) = E(X-A)^{\otimes l} \underset{(2.12)}{=} E\Big[\sum_{r=0}^{l}\binom{l}{r} X^{\otimes r}\otimes(-A)^{\otimes l-r}\Big] =$$

$$= \sum_{r=0}^{l}\binom{l}{r}(-A)^{\otimes l-r}\otimes EX^{\otimes r} = \sum_{r=0}^{l}\binom{l}{r}(-1)^{\otimes l-r}A^{\otimes l-r}\otimes m^{(r)}(X),$$

particularly, for $A = EX = m^{(1)}(X) = m^{(1)}$, we have

$$\mu^{(1)} = m^{(1)} - A = 0,$$

$$\mu^{(2)} = m^{(2)} - 2A\otimes m^{(1)} + A^{\otimes 2} = m^{(2)} - (m^{(1)})^{\otimes 2},$$

$$\mu^{(3)} = m^{(3)} - 3A\otimes m^{(2)} + 3A^{\otimes 2}\otimes m^{(1)} - A^{\otimes 3} =$$

$$= m^{(3)} - 3m^{(2)}\otimes m^{(1)} + 2(m^{(1)})^{\otimes 3}.$$

6.4. Let the elements x and y of random arrays X and Y be independent random variables for every pair (x,y). Whereas Exy = ExEy for the independent random variables x and y, we get for the random arrays X and Y

a) $I \cap J = \emptyset,\ \exists EX_{IK}, EY_{KJ} \to \exists E(X_{IK}Y_{KJ}),$

$$E(X_{IK}Y_{KJ}) = [E\sum_{k\in K} x_{ik}y_{kj}] = [\sum_{k\in K} Ex_{ik}y_{kj}] = (EX_{IK})(EY_{KJ}),$$

b) $\exists EX_I,\ EY_J,\ (X_I\otimes Y_J) \to \exists E(X_I\otimes Y_J),$

$$E(X_I\otimes Y_J) = [E(x_iy_j)] = [Ex_i\cdot Ey_j] = EX_I\otimes EY_J.$$

Therefore from (6.1) follows that

a) $I\cap J = \emptyset,\ \exists E|x_{ik}|^l,\ E|y_{kj}|^l\ \forall i\in I,\ k\in K,\ j\in J \to$

$$\to \exists m^{(l)}(X_{IK}Y_{KJ}),\ m^{(l)}(X_{IK}Y_{KJ}) = E(X_{IK}Y_{KJ})^{\otimes l} \underset{(2.11)}{=}$$

$$= E(X^{\otimes l}_{IK}Y^{\otimes l}_{KJ}) = (EX^{\otimes l}_{IK})(EY^{\otimes l}_{KJ}) = m^{(l)}(X_{IK})m^{(l)}(Y_{KJ}),$$

b) $\exists(X_I\otimes Y)_J,\ \exists E|x_i|^l,\ E|y_j|^l\ \forall i\in I,\ j\in J \to$

$$\to \exists m^{(l)}(X_I \otimes Y_J),$$

$$m^{(l)}(X_I \otimes Y_J) = E(X_I \otimes Y_J)^{\otimes l} \underset{(2.10)}{=} E(X_I^{\otimes l} \otimes Y_J^{\otimes l}) =$$

$$= (EX_I^{\otimes l}) \otimes (EY_J^{\otimes l}) = m^{(l)}(X_I) m^{(l)}(Y_J),$$

c) $\exists E|x_i|^l$, $E|y_j|^l$ $\forall i \in I \to \exists m_{A+B}^{(l)}(X+Y)$, $m_{A+B}^{(l)}(X+Y) =$

$$= \sum_{r=0}^{l} \binom{l}{r} (E(X-A)^{\otimes r}) \otimes (E(Y-B)^{\otimes l-r}) = \sum_{r=0}^{l} \binom{l}{r} m_A^r(X) m_B^{\otimes l-r}(Y).$$

7. Characteristic function of the random array

7.1. For the real random array X_J we define the characteristic function φ as the complex-valued function of the real array T_J as follows:

$$\varphi(T_J) = \varphi(X_J; T_J) = Ee^{iT_J X_J} \text{ with } i = \sqrt{-1}. \tag{7.1}$$

(We note that $\varphi(T_J)$ exists for every T_J whereas $|\mathrm{Re}\, e^{iT_J X_J}| \le 1$, $|\mathrm{Im}\, e^{iT_J X_J}| \le 1$.)

7.2. We recall the following property of the mean of the random variable and the random vector X:

let $g(X,t)$ be a function so that there exist $Eg(X,t)$ and $E\left|\frac{d}{dt}g(X,t)\right|$; then there exist and are equal $E\frac{d}{dt}g(X,t)$ and $\frac{d}{dt}Eg(X,t)$.

Using this result, it is not difficult to prove (by induction by r) the following Theorem 7.1.

Theorem 7.1. Let $E|x_j|^l < \infty$ for all the elements of the real random array X_J. Then the r-th array-derivative of φ by T_J exists for every $r = 0,1,\ldots,l$ and

$$\frac{d^r \varphi(X_J; T_J)}{dT_J^r} = E(i^r X_J^{\otimes r} e^{iT_J X_J}), \tag{7.2}$$

$$\left.\frac{d^r \varphi(X_J; T_J)}{dT_J^r}\right|_{T_J = 0_J} = i^r EX_J^{\otimes r} = i^r m^{(r)}(X). \tag{7.3}$$

Proof. For $l = 0$ we have $\frac{d^0 \varphi(X;T)}{dT^0} = \varphi(X;T) = Ee^{iT_J X_J}$,

$$\left.\frac{d^0 \varphi(X;T)}{dT^0}\right|_{T=0} = \left. Ee^{iT_J X_J}\right|_{T=0} = E1 = 1 = i^0 m^{(0)}(X).$$

Let (7.2) hold for $r = 0,1,\ldots,s < 1$. Then

$$\frac{d^{s+1}\varphi(X;T)}{dT^{s+1}} = \frac{d}{dT}\left(\frac{d^{s}\varphi(X;T)}{dT^{s}}\right) = \frac{d}{dT}E(i^{s}X_{j}^{\otimes s}e^{iT_{j}X_{j}}).$$

The means of $i^{s}x_{j_1}\cdot\ldots\cdot x_{j_s}e^{iT_{j}X_{j}}$ and of

$$\left|\frac{\partial}{\partial t_{j_{s+1}}}(i^{s}x_{j_1}\cdot\ldots\cdot x_{j_s}e^{iT_{j}X_{j}})\right| =$$

$$= \left|i^{s+1}x_{j_1}\cdot\ldots\cdot x_{j_{s+1}}e^{iT_{j}X_{j}}\right| = \left|x_{j_1}\cdot\ldots\cdot x_{j_{s+1}}\right|$$

are finite. Therefore the elements $\frac{\partial}{\partial t_{j_{s+1}}}E(i^{s}x_{j_1}\cdot\ldots\cdot x_{j_s}e^{iT_{j}X_{j}})$ of the array $\frac{d}{dT}E(i^{s}X_{j}^{\otimes s}e^{iT_{j}X_{j}})$ and the elements

$$E\frac{\partial}{\partial t_{j_{s+1}}}(i^{s}x_{j_1}\cdot\ldots\cdot x_{j_s}e^{iT_{j}X_{j}}) = i^{s+1}x_{j_1}\cdot\ldots\cdot x_{j_{s+1}}e^{iT_{j}X_{j}}$$

of the array $E(i^{s+1}X_{j}^{\otimes s+1}e^{iT_{j}X_{j}})$ exist and are equal for all $j_1,\ldots,j_{s+1} \in J$. Thus (7.2) holds and therefore (7.3) also holds.

Acknowledgement. The author is grateful to Dr. T. Kollo, who inspired to carry out this work.

References

Kollo, T. (1991). Matrix Derivative in Multivariate Statistics. Tartu University Press, Tartu (in Russian).

Magnus, J. R., Neudecker, H. (1988). Matrix Differential Calculus with Applications in Statistics and Economics. Wiley, Chichester.

Pollock, D. S. G. (1985). Tensor products and matrix differential calculus. Linear Algebra and its Applications, 67, 169-193.

Rogers, G. S. (1980). Matrix Derivatives. Marcel Dekker, New York.

CONTENTS