Arnold, Matthias; Weißbach, Rafael

Working Paper

Testing large-dimensional correlation

Technical Report, No. 2007,15

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Arnold, Matthias; Weißbach, Rafael (2007) : Testing large-dimensional correlation, Technical Report, No. 2007,15, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/25000

Testing Large-Dimensional Correlation

Matthias Arnold & Rafael Weißbach*

Institute of Business and Social Statistics, University of Dortmund, Dortmund, Germany

2nd May 2007

Abstract

This paper introduces a test for zero correlation in situations where the correlation matrix is large compared to the sample size. The test statistic is the sum of the squared correlation coefficients in the sample. We derive its limiting null distribution as the number of variables as well as the sample size converge to infinity. A Monte Carlo simulation finds both size and power for finite samples to be suitable. We apply the test to the vector of default rates, a risk factor in portfolio credit risk, in different sectors of the German economy.

Keywords. Testing correlation, *n-p-* asymptotics, portfolio credit risk.

*Corresponding author: Rafael Weißbach, Institute of Business and Social Statistics, Faculty of Statistics, University of Dortmund, 44221 Dortmund, Germany, email: Rafael.Weissbach@uni-dortmund.de, Fon: +49/231/7555419, Fax: +49/231/7555284.
JEL subject classifications. C12, C52.

1 Introduction

Many applications of multivariate analysis involve a large number of variables. However, the associated statistical procedures are often derived from large sample asymptotics. In practice, it may occur that the sample size is smaller than the number of variables, a situation which is often ruled out in testing for correlation, for instance for the likelihood ratio test (see Muirhead (1982), p. 527 or Anderson (1958), p. 233).

When the number of variables is of the same order of magnitude as the sample size, the finite performance of such procedures is doubtful. Therefore, one should rather use procedures that are based on an asymptotic theory in which both sample size and the number of variables converge to infinity. Examples of such procedures can be found in Dempster (1958), Ledoit and Wolf (2002) or Schott (2005).

In this paper, we consider a correlation test for p normally distributed random variables. We analyse a test statistic which works when the dimensionality is large: the sum of squared correlation coefficients in the sample. This statistic is also considered by Schott (2005), who shows that its distribution is asymptotically normal as sample size and number of variables both tend to infinity. We analyse the null distribution of this statistic in three asymptotic situations. Using results about the asymptotic joint distribution of sample correlation coefficients by Browne and Shapiro (1986) and Neudecker and Wesselmann (1990), we show that the distribution of the test statistic converges to a χ^2-distribution as the sample size converges to infinity. Starting from that basis, we approximate the distribution for a fixed sample size as the dimension increases. Finally, we consider the asymptotic distribution for the case in which both the sample size and the dimension converge to infinity. It turns out that this n-p-asymptotic result is the same as for the fixed dimension. Simulations show that our test has sufficient power and keeps the nominal size for reasonable sample sizes.

We apply the test to a high-dimensional risk factor for a credit portfolio

model. Typically, few current data are available to estimate the distribution of the risk factor, and this is specially critical for the correlation matrix due to the "curse of dimensionality". The specific example is the vector of default rates in different sectors of the economy. It serves as a risk factor in e.g. the model CreditRisk$^+$ (Credit Suisse First Boston (CSFB) (1997)). The correlations between the default rates enter into the credit value-at-risk of a loan portfolio (Bürgisser et al. (1999)). For our data, the null hypothesis of no intersectional correlation is clearly rejected.

Moreover, we apply the test to a data set containing 8 blood serum measurements for a group of 12 individuals, made up of 4 alcoholics and 8 controls. For each of the two subgroups, we test whether the different blood serum measurements are correlated or not. The data exhibit very strong evidence of correlation for the control group, whereas for the alcoholic group there is less evidence for correlation.

2 Model and main results

A frequent assumption in multivariate statistics is that of an i.i.d. sample of normally distributed p-dimensional random vectors $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$, in which $\mu \in \mathbb{R}^p$ denotes the mean vector and $\Sigma \in \mathbb{R}^{p \times p}$ a positive semi-definite covariance matrix. The corresponding correlation matrix is $P = \Sigma_0^{-\frac{1}{2}} \Sigma \Sigma_0^{-\frac{1}{2}}$, where Σ_0 is a diagonal matrix with the same diagonal elements as Σ. The usual estimator for P is the sample correlation matrix R with typical entries

$$r_{ij} = \frac{\sum_{k=1}^n (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_{k=1}^n (X_{ik} - \bar{X}_i)^2 \sum_{k=1}^n (X_{jk} - \bar{X}_j)^2}}.$$

Here X_{ik} is the i^{th} element of the variable X_k, $\bar{X}_i$ is the mean. We consider the following hypothesis

$$H_0 : P = I \text{ vs. } H_1 : P \neq I,$$

where I denotes the $p \times p$ unit matrix. H_0 corresponds to the independence of the coordinates. The coordinates' variances are nuisance parameters, as their expectations.

For the related testing problem $H_0^{'} : \Sigma = I$, John (1971) suggested the statistic

$$tr(S-I)^2 = \sum_{i=1}^{p} \sum_{j=1, j \neq i}^{p} s_{ij}^2 + \sum_{i=1}^{p} (s_{ii}-1)^2, \tag{1}$$

where $tr(\cdot)$ denotes the trace and S is the usual unbiased estimator for Σ. If $\Sigma = I$, $\sigma_{ii} = 1$ and $\sigma_{ij} = 0$ for $i \neq j$, then (1) is clearly a reasonable statistic. It can also be used when $n < p$.

Similarly, we use the statistic $tr(R-I)^2$ for the correlation test. Because R is symmetric, the statistic equals twice the sum of the squared correlation coefficients in the sample:

$$\begin{aligned} tr(R-I)^2 &= tr\left(diag\left(\sum_{i=2}^{p} r_{1i}^2, \sum_{i=1, i \neq 2}^{p} r_{2i}^2, \ldots, \sum_{i=1}^{p-1} r_{pi}^2 \right) \right) \\ &= \sum_{i=1}^{p} \sum_{j=1, j \neq i}^{p} r_{ij}^2 \\ &= 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} r_{ij}^2. \end{aligned} \tag{2}$$

Under H_0, the r_{ij}'s should be close to zero. We will reject H_0 if (2) is too large and determine the distribution under H_0 with the aid of an asymptotic approximation. Since asymptotics for a fixed p and $n \to \infty$ (n-asymptotics) contradicts the situation considered here, an asymptotic approximation where both $n \to \infty$ and $p \to \infty$ is used. This will be done in three stages. First, we derive the asymptotic distribution for large samples when the dimension is fixed. Next, the sample size is fixed and the dimension increases. Finally, we combine the results of the previous steps to derive the asymptotic distribution when the sample size and the dimension both converge to infinity.

Lemma 2.1 gives the joint asymptotic distribution of the correlation coefficients under H_0, for $n \to \infty$.

Lemma 2.1 *Let $X_1, \ldots, X_n$ be i.i.d. $\sim N_p(\mu, \Sigma)$ with $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. Then for $n \to \infty$,*

$$\sqrt{n-1}\ (r_{12}, r_{13}, \ldots, r_{p-1,p})^t \xrightarrow{\mathcal{D}} N_{p(p-1)/2}\left(0, I_{p(p-1)/2}\right).$$

Proof. Browne and Shapiro (1986) show that for arbitrary p,

$$\sqrt{n-1}(r_{11}, r_{12}, \ldots, r_{p-1,p}, r_{p,p})^t$$

approaches a p^2-variate normal distribution with expectation

$$\sqrt{n-1}(\rho_{11}, \rho_{12}, \ldots, \rho_{p-1,p}, \rho_{p,p})^t.$$

The asymptotic covariance matrix is given by

$$2M_s(P \otimes P) - AB^t - BA^t + AGA^t. \tag{3}$$

Here, $\otimes$ denotes the Kronecker symbol. In the following, double subscripts are used to denote rows of a matrix with p^2 rows, or columns of a matrix with p^2 columns. Thus, $C_{ij,kl}$ represents the element in row $(j-1)p+i$ and column $(l-1)p+k$ of a matrix C. The matrices in (3) have typical elements $(2M_s(P \otimes P))_{ij,kl} = \rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk}$, $(B)_{ij,k} = 2\rho_{ik}\rho_{jk}$ and $(G)_{ij} = 2\rho_{ij}^2$. A is defined as $A = M_s(I \otimes P)K_d$ where the only nonzero elements of M_s and K_d are

$$(M_s)_{ij,ij} = (M_s)_{ij,ji} = \begin{cases} 1 & \text{if } i = j \\ \frac{1}{2} & \text{if } i \neq j \end{cases}$$

and $(K_d)_{ii,i} = 1$. Now, the three-dimensional case is considered first. Accord-

ingly, $P = I_3$, and

$$2M_s(P \otimes P) = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}, A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$B = 2A$ and $G = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ so that

$$2M_s(P \otimes P) - AB^t - BA^t + AGA^t = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

With respect to rows and columns 2, 3 and 6 only, the asymptotic covariance matrix of $\sqrt{n-1}(r_{12}, r_{13}, r_{23})^t$ turns out to be I_3.

This special case of $p = 3$ can easily be generalized to an arbitrary p: since the asymptotic variance of each r_{ij} is clearly the same, it remains to show that r_{ij} and r_{kl} are asymptotically uncorrelated, provided at least one of the indices is different. If both indices are different, r_{ij} and r_{kl} are independent, even for a finite n. If only one index is different, they are asymptotically

independent, because the covariance between r_{ij} and r_{ik} for $j \neq k$ is the same as the covariance between r_{12} and r_{13}, which has been shown to be asymptotically negligible. □

Lemma 2.1 immediately yields the n-asymptotic null-distribution of the test statistic expressed in Theorem 2.2.

Theorem 2.2 *Let $X_1, \ldots, X_n$ be i.i.d. $\sim N_p(\mu, \Sigma)$ with $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. Then for $n \to \infty$,*

$$\frac{n-1}{2}\, tr(R-I)^2 \xrightarrow{\mathcal{D}} \chi^2_{p(p-1)/2}. \tag{4}$$

Proof. The proposition follows directly from (2), Lemma 2.1 and the definition of the χ^2-distribution. □

The test procedure using the n-asymptotic approximation is

$$\phi_n(X_1, \ldots, X_n) = \mathbf{1}_{\left\{\frac{n-1}{2}\ tr(R-I)^2 > \chi^2_{p(p-1)/2,1-\alpha}\right\}}$$

where $\mathbf{1}_{\{\}}$ denotes the indicator function and $\chi^2_{p(p-1)/2,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with $p(p-1)/2$ degrees of freedom.

Our next result gives the asymptotic approximation for a fixed sample size n, as the dimension p converges to infinity. In this situation, the number of squared correlation coefficients contained in $tr(R-I)^2$ increases. Because each r_{ij}^2 follows a beta distribution with parameters $1/2$ and $(n-2)/2$ (see Muirhead (1982), p. 147), for $i \neq j$, we have

$$\mathrm{E}(r_{ij}^2) = \frac{1}{n-1}, \quad \mathrm{Var}(r_{ij}^2) = \frac{2(n-2)}{(n-1)^2(n+1)}. \tag{5}$$

Two different squared correlation coefficients r_{ij}^2 and r_{kl}^2 are now examined. If all indices are different, these are independent for all n. If one index is the same, they are dependent, but uncorrelated for a fixed n. For a very large n, they tend to be independent, because of the asymptotic normality of r_{ij}

and r_{kl}. A proof of the lack of correlation of the corresponding covariance estimators s_{ij} is set out in Lemma A.1 in the appendix. Lemma 2.3 gives the asymptotic distribution of a sum of independent random variables, each of which follows a beta distribution.

Lemma 2.3 *Let* $Y_{12}, Y_{13} \ldots, Y_{p-1,p}$ *be i.i.d* $\sim Beta(1/2, (n-2)/2)$. *Then for* $p \to \infty$,

$$(n-1) \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} Y_{ij} \sqrt{\frac{n+1}{p(p-1)(n-2)}} - \sqrt{\frac{p(p-1)(n+1)}{4(n-2)}} \xrightarrow{\mathcal{D}} N(0,1). \quad (6)$$

Proof. The proposition follows from (5) and the central limit theorem. □

Since Lemma 2.3 requires that the random variables be independent, it is applicable to the sum of squared correlation coefficients for an infinite sample size. We combine the preceding asymptotic approximations with Lemma 2.3, in order to derive the distribution of $(n-1)tr(R-I)^2/2$, when n and p both converge to infinity.

Theorem 2.4 *Let* $X_1, \ldots, X_n$ *be i.i.d.* $\sim N_p(\mu, \Sigma)$ *with* $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. *Then for* $n, p \to \infty$,

$$\frac{n-1}{2}\, tr(R-I)^2 \sqrt{\frac{n+1}{n-2}} - \frac{p(p-1)}{2} \left(\sqrt{\frac{n+1}{n-2}} - 1 \right) \xrightarrow{\mathcal{D}} \chi^2_{p(p-1)/2}. \quad (7)$$

Proof. For $n \to \infty$, $\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} Y_{ij}$ in (6) can be replaced by $tr(R-I)^2/2$, because the r_{ij}^2's are independent in this case. This expression approaches to $N(0,1)$. Note that a sum of $p(p-1)/2$ independent random variables Y_k, each of which follows a χ_1^2-distribution, has a $\chi^2_{p(p-1)/2}$-distribution. Thus, with $\mathrm{E}(Y_k) = 1$, $\mathrm{Var}(Y_k) = 2$, the central limit theorem yields

$$\frac{1}{\sqrt{p(p-1)}} \chi^2_{p(p-1)/2} - \sqrt{\frac{p(p-1)}{4}} \xrightarrow{p \to \infty} N(0,1). \quad (8)$$

Since the left hand sides of (6) and (8) both tend towards the same limiting distribution, the left hand side of (6) is distributed according to the left hand side of (8), so that

$$\begin{aligned} \tfrac{n-1}{2}\, tr(R-I)^2 \quad & \sqrt{\frac{n+1}{p(p-1)(n-2)}} - \sqrt{\frac{p(p-1)(n+1)}{4(n-2)}} \\ & \stackrel{n,p\to\infty}{\longrightarrow} \frac{1}{\sqrt{p(p-1)}}\chi^2_{p(p-1)/2} - \sqrt{\frac{p(p-1)}{4}}. \end{aligned}$$

Multiplying both sides by $\sqrt{p(p-1)}$ and adding $p(p-1)/2$ completes the proof. □

Using the n-p-asymptotic approximation (7), we obtain our final test procedure

$$\phi_{np}(X_1,\ldots,X_n) = \mathbf{1}_{\left\{\frac{n-1}{2}\, tr(R-I)^2\sqrt{\frac{n+1}{n-2}}-\frac{p(p-1)}{2}\ \sqrt{\frac{n+1}{n-2}}-1\ >\chi^2_{p(p-1)/2,1-\alpha}\right\}}.$$

Remark 2.5 For large n, $\sqrt{(n+1)/(n-2)}$ is near to one, so that the n-p-asymptotic approximation (7) finally corresponds to the n-asymptotic approximation (4). Calculating expectation and variance of $(n-1)tr(R-I)^2/2$ as

$$\begin{aligned} \mathrm{E}\left(\frac{n-1}{2}\, tr(R-I)^2\right) &= (n-1)\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{1}{n-1} = \frac{p(p-1)}{2}, \\ \mathrm{Var}\left(\frac{n-1}{2}\, tr(R-I)^2\right) &= (n-1)^2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{2(n-2)}{(n-1)^2(n+1)} \\ &= p(p-1)\frac{n-2}{n+1}, \end{aligned}$$

we see that (7) provides a finite adjustment for (4). The expectation of $(n-1)tr(R-I)^2/2$ conforms to the $\chi^2_{p(p-1)/2}$-distribution, but the variance is smaller than $p(p-1)$ for a finite n. Therefore, the statistic is multiplied by a constant greater than one. Thus the expectation becomes too large. Therefore, as an adjustment, a second constant must be subtracted.

Table 1: Actual size of the test (nominal size $\alpha = 0.05$)

$p\backslash n$		5	10	20	40	100
5	ϕ_n	0.02	0.03	0.04	0.04	0.05
	ϕ_{np}	0.05	0.05	0.05	0.05	0.05
10	ϕ_n	0.02	0.03	0.04	0.04	0.05
	ϕ_{np}	0.06	0.05	0.05	0.05	0.05
20	ϕ_n	0.02	0.04	0.04	0.05	0.05
	ϕ_{np}	0.06	0.05	0.05	0.05	0.05
40	ϕ_n	0.02	0.04	0.04	0.05	0.05
	ϕ_{np}	0.07	0.06	0.05	0.05	0.05
100	ϕ_n	0.03	0.04	0.04	0.05	0.05
	ϕ_{np}	0.07	0.06	0.05	0.05	0.05

3 Some Monte Carlo simulations

We draw n-sized samples from a p-variate normal distribution with expectation $0 \in \mathbb{R}^p$ and covariance matrix Σ. We choose $\Sigma = I$ for the size simulations and

$$\Sigma = \begin{pmatrix} 1 & 0.2 & \dots & 0.2 \\ 0.2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.2 \\ 0.2 & \dots & 0.2 & 1 \end{pmatrix}$$

for the power simulations. Because correlation coefficients are invariant with respect to transformations of location and scale, this approach is reliable enough as far as the size is concerned. With respect to the power, we restrict ourselves to the case of equicorrelation with $\rho = 0.2$. For each combination of n and p in 5, 10, 20, 40 and 100, we use 10,000 repetitions. Critical values correspond to a nominal size of 0.05. Tables 1 and 2 show the empirical

Table 2: Power of the test (nominal size $\alpha = 0.05$)

$p\backslash n$		5	10	20	40	100
5	ϕ_n	0.04	0.14	0.33	0.67	0.98
	ϕ_{np}	0.09	0.19	0.37	0.68	0.98
10	ϕ_n	0.08	0.30	0.66	0.95	1.00
	ϕ_{np}	0.17	0.35	0.67	0.95	1.00
20	ϕ_n	0.19	0.54	0.90	1.00	1.00
	ϕ_{np}	0.29	0.60	0.91	1.00	1.00
40	ϕ_n	0.35	0.77	0.99	1.00	1.00
	ϕ_{np}	0.43	0.80	0.99	1.00	1.00
100	ϕ_n	0.59	0.94	1.00	1.00	1.00
	ϕ_{np}	0.66	0.95	1.00	1.00	1.00

rejection frequencies.

In Table 1, we see that the actual size of the tests is close to the nominal size of 0.05 for all combinations of n and p. As expected, the n-asymptotic approximation ϕ_n exhibits downward size distortion for a small n. The actual size of ϕ_{np} is closer to the nominal one. Upward distortions of ϕ_{np} for small n and large p seem to be minimal. These empirical findings are in line with analytic results.

The simulated power in Table 2 shows that both tests are n-consistent and n-p-consistent. Even for a fixed sample size, a larger dimension leads to rapidly increasing power. The power of the two tests does not differ much, and the differences seem to be caused by the varying actual sizes. This is due to the fact that both test procedures rely on the same statistic $\mathrm{tr}(R - I)^2$, only the standardizations differ.

Consequently, we recommend ϕ_{np} because its approximation appears more accurate than that of ϕ_n.

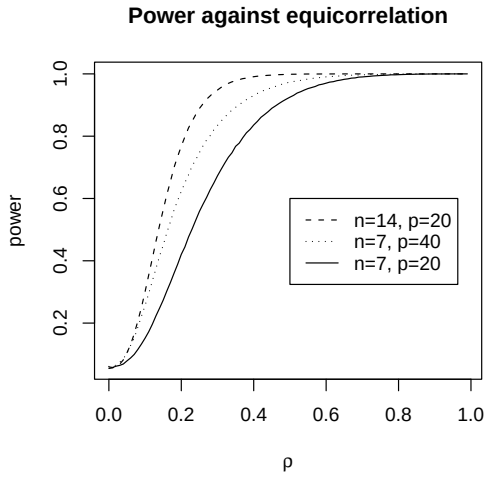

Figure 1: Empirical rejection probabilities of ϕ_{np} for the case of equicorrelation

Figure 1 displays empirical rejection probabilities of the test ϕ_{np} for the case of equicorrelation with different values of ρ. The case where ρ is equal to 0 stands for the null hypothesis, whereas positive values of ρ represent different points of the alternative.

The combination of n equal to 7 and p equal to 20 refers to the situation of our data set containing default rates. As expected, the solid line shows that rejection probabilities rise if ρ gets larger. Figure 1 also exhibits how the power increases if the number of variables (dotted line) or the sample size (dashed line) is doubled. The effect of a rising sample size is more distinct than the effect of a growing number of variables.

4 Empirical examples

We apply our test procedure ϕ_{np} to two data sets. The first one contains default rates in different sectors of the German economy. The data covers 20 sectors over a period of 7 years. The data set is a typical example of large-dimensionality: on the one hand, there should be a substantial number of sectors included because e.g. banks use very detailed sector classifications for credit customers. On the other hand, the data should not be too old, because the amount of information contained in the data tends to decrease over time.

Using ϕ_{np} to test for correlation between the default rates in different sectors, we calculate (7) and obtain a value for the test statistic of 1190, corresponding to a p-value of virtually zero. The null hypothesis of no intersectional correlation is clearly rejected.

This result agrees with a priori expectations: after all, different sectors of the economy depend on the same macroeconomic variables.

The second data are part of a larger data set in Beerstecher Jr. et al. (1950). It consists of 8 blood serum measurements for a group of 12 individuals, made up of 4 alcoholics and 8 controls. For each of the two subgroups, we test whether the different blood serum measurements are correlated or not. Again we use ϕ_{np} and calculate the value of the test statistic as 62.0 for the control group and 39.7 for the group of alcoholics, corresponding to p-values of 0.0002 and 0.0696, respectively. Thus the data exhibit very strong evidence of correlation for the control group, whereas for the alcoholic group there is less evidence for correlation.

Acknowledgements: We are indebted to Walter Krämer and Götz Trenkler for valuable comments. Furthermore, the financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures", project B1) is gratefully acknowledged.

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, 1958.

E. Beerstecher Jr., E. Sutton, H.K. Berry, W.D. Brown, J. Reed, G.B. Rich, L.J. Berry, and R.J. Williams. Biochemical individuality. v. explorations with respect to metabolic patterns of compulsive drinkers. *Archives of Biochemistry*, 29:27–40, 1950.

M. W. Browne and A. Shapiro. The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra and its Applications*, 82:169–176, 1986.

P. Bürgisser, A. Kurth, A. Wagner, and M. Wolf. Integrating correlations. *Risk magazine*, 12:57–60, 1999.

Credit Suisse First Boston (CSFB). CreditRisk$^+$: A credit risk management framework. Technical report, Credit Suisse First Boston, 1997.

A. P. Dempster. A high dimensional two sample significance test. *Annals of Mathematical Statistics*, 29:995–1010, 1958.

S. John. Some optimal multivariate tests. *Biometrika*, 58:123–127, 1971.

O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, 30:1081–1102, 2002.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, New York, 1982.

H. Neudecker and A. M. Wesselmann. The asymptotic variance matrix of the sample correlation matrix. *Linear Algebra and its Applications*, 127: 589–599, 1990.

J. R. Schott. Testing for complete independence in high dimensions. *Biometrika*, 92:951–956, 2005.

A Correlation between correlation coefficients

Lemma A.1 *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n, Z_1, \ldots, Z_n$ be independently distributed with expectation μ_X, μ_Y, μ_Z and variances $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$, respectively. Then, the covariance estimators*

$$s_{XY} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right),$$

s_{XZ} and s_{YZ} satisfy

$$Cov(s_{XY}, s_{XZ}) = Cov(s_{XY}, s_{YZ}) = Cov(s_{XZ}, s_{YZ}) = 0.$$

Proof. Let 1_n be the n-dimensional vector with all elements equal to one and $H = I - 1/n 1_n 1_n^t$. With $X = (X_1, \ldots, X_n)^t$, $Y = (Y_1, \ldots, Y_n)^t$ and $Z = (Z_1, \ldots, Z_n)^t$,

$$s_{XY} = \frac{1}{n-1}X^t H Y, \quad s_{XZ} = \frac{1}{n-1}X^t H Z, \quad s_{YZ} = \frac{1}{n-1}Y^t H Z.$$

It follows that apart from the constant factor $1/(n-1)^2$,

$$\begin{aligned}
\mathrm{Cov}(s_{XY}, s_{XZ}) &= \mathrm{E}(s_{XY} \cdot s_{XZ}) - \mathrm{E}(s_{XY}) \cdot \mathrm{E}(s_{XZ}) \\
&= \mathrm{E}(X^t H Y X^t H Z) - \mathrm{E}(X^t H Y) \cdot \mathrm{E}(X^t H Z) \\
&= \left[\mathrm{E}(X^t H Y X^T) H \mathrm{E}(Z) - \mathrm{E}(X^t) H \mathrm{E}(Y) \mathrm{E}(X^t) H \mathrm{E}(Z)\right] \\
&= \left[\mathrm{E}(X^t H Y X^t) H \mu_Z 1_n - \mathrm{E}(X^t) H \mu_Y 1_n \mathrm{E}(X^t) H \mu_Z 1_n\right] \\
&= 0,
\end{aligned}$$

since $H 1_n = 0$. □