Sankhyā: The Indian Journal of Statistics
1998, Volume 60, Series A, Pt. 1, pp. 119-133

SPHERING AND ITS PROPERTIES

By GUOYING LI
and
JIAN ZHANG
Academia Sinica, Beijing

SUMMARY. Centering and sphering is an intuitive approach to remove, in a sense, location, scale and correlation structure in data sets and to force us to examine other aspects of them. It is frequently applied in data analyses. This paper is intended to discuss properties of sphering procedures, such as affine (including orthogonal and lower triangular as special cases) equivariance/invariance, application to projection pursuit (PP) and asymptotic behavior. In particular, the three commonly used sphering procedures, named LTS, SRS and JFS, are studied. It is shown that all sphering methods in a PP-after-sphering procedure results in the same optimal projections. It is also shown that the sphering matrix of JFS is inconsistent, whereas, those of SRS and LTS not only are consistent but also have asymptotic distributions.

1. **Introduction**

In classical multivariate analysis the sample mean and covariance matrix are the most frequently applied summary statistics. However, they can mainly capture linear correlation or elliptic structures, which are now well-understood features in elliptic shaped data sets. For many purposes of graphical displays and other exploratory analyses of high dimensional data sets, one wants to look beyond linear correlation or elliptic structures. Outliers, clusters or other kind of groups, and concentrations near curves or non-flat surfaces are probably the important features that interest data analysts. They are, in general, not obtainable through mere knowledge of the sample mean and covariance matrix. In these circumstances, it is desirable to separate off the information contained in the mean and the covariance matrices and forces us to examine aspects of our data sets other than those well-understood natures. Centering and sphering is a

---

Paper received. February 1995; revised November 1997.
AMS (1991) *subject classification.* Primary 62H99.
Key words and phrases. Sphering, equivariance, invariance, asymptotics, projection pursuit.

simple and intuitive approach that eliminates the mean-covariance information and helps to highlight structures beyond linear correlation and elliptic shapes, and therefore is often performed before exploring displays or analyses of data sets.

In order to reach the same goal in exploring projection pursuit (PP for short) analyses, an affine invariant projection index is required (see Huber 1985, Section 5.2). If such a projection index is not readily available, the analysts can, instead, simply perform a centering and sphering transformation before doing projection pursuit (see Subsection 2.2 below). Another advantage of centering and sphering is a substantial computational saving in exploring PP (see Friedman 1987).

The role of centering and sphering was discussed by Tukey and Tukey (1981), Huber (1981), Friedman (1987), and Jones and Sibson (1987). The last three papers are all concerned with PP techniques.

Centering and sphering can be specified as follows. Let $x_1, \cdots, x_n$ be $p-$vectors observed from population x with location μ and positive definite dispersion matrix Σ. Let $X_n = (x_1, \cdots, x_n)$ or $X = (x_1, \cdots, x_n)$ (the subscript n is usually omitted for fixed sample size n) which is a $p \times n$ matrix called data matrix. Let $\mu_n = \hat{\mu}(X_n)$ and $\Sigma_n = \hat{\Sigma}(X_n)$ be estimators of μ and Σ based on X_n, respectively. Suppose Σ_n is positive definite. Centering and sphering is a transformation of the form

$$x_i \longrightarrow B_n(x_i - \mu_n) \quad \text{with} \quad B_n^\tau B_n = \Sigma_n^{-1}, \qquad \ldots(1.1)$$

where B_n is a $p \times p$ matrix. Clearly, the transformed data has zero location and unit dispersion matrix. We call B_n a *sphering matrix* for the data set X_n because of its sphering effect. Obviously, B_n is not uniquely determined by Σ_n and the requirement in (1.1). It can be obtained by many decomposition methods.

Thus, sphering procedure consists of two parts: an estimator Σ_n of dispersion matrix, and a decomposition method to determine B_n. Estimation of dispersion matrix is extensively investigated in literature. There are various estimators with different properties. Tukey and Tukey (1981) briefly discussed the choices of dispersion estimators.

There are three decomposition methods to determine sphering matrices that appear in literature: lower triangular decomposition (LTD), square root decomposition (SRD) and one adopted by Friedman (1987) named JFD here. This paper is intended to study properties of sphering. In Section 2, we first specifically formalize sphering and related concepts. Then properties of sphering transformations, especially those based on the three decomposition methods mentioned above, are studied, including equivariance/invariance, their effects when applied to projection pursuit and the asymptotic behavior of the sphering matrices obtained by LTD, SRD and JFD respectively. The technical argument is presented in Section 4. Section 3 is devoted to a discussion of related topics.

2. Properties of Sphering

We consider only affine equivariant estimators for location and dispersion matrix, i.e., for any nonsingular A and $p-$vector b,

$$\begin{aligned}\hat{\mu}(Ax_1+b,\cdots,Ax_n+b) &= A\hat{\mu}(X)+b\\ \hat{\Sigma}(Ax_1+b,\cdots,Ax_n+b) &= A\hat{\Sigma}(X)A^{\tau}.\end{aligned}$$

Let B_n be a $p\times p$ matrix and $\Sigma_n^{-1}=B_n^{\tau}B_n$ be a certain decomposition, say decomposition $*$, of Σ_n^{-1}. Denote by $\mathbf{1}=(1,\cdots,1)^{\tau}$ a $p-$vector whose elements are all $1'$s. Then, centering and sphering is, as mentioned in previous section, a map defined by $CS(X_n)=B_n(X_n-\mu_n\mathbf{1}^{\tau})$.

In practice, one always spheres centered data matrices. From the translation invariance of dispersion estimators $\hat{\Sigma}(\cdot)$, we see that the sphering matrices have nothing to do with centering. Also, from the affine equivariance of $\hat{\mu}(\cdot)$, it follows that $AX_n-\hat{\mu}(AX_n)=A[X_n-\hat{\mu}(X_n)]$, for any $p\times p$ matrix A. That is, affine transformation and centering commute. Hence, without loss of generality, we can assume in this paper that $\hat{\mu}(X_n)=0$, i.e., the data matrices are already centered. This allows us to discuss the "pure" sphering transformation (or simply sphering), i.e., a map $S(\cdot)$ defined by $S(X_n)=B_nX_n$, where B_n is as stated in (1.1) and obtained by decomposition $*$. Now, let us formally define the three commonly used spherings mentioned in Section 1.

Let $\Sigma_n=A(n)D(n)A^{\tau}(n)$ be a spectral decomposition of Σ_n, where $A(n)$ is an orthogonal matrix and $D(n)=diag(d(n,1),\cdots,d(n,p))$ with $d(n,1)\geq\cdots\geq d(n,p)$. Let $\Sigma_n^{-1/2}$ and $D(n)^{-1/2}$ be defined as usual, i.e., for $\Lambda=diag(\lambda_1,\cdots,\lambda_r,0,\cdots,0)$ with $\lambda_i>0$, $i=1,\cdots,r$, and orthogonal matrix P,

$$\Lambda^{\kappa}=diag(\lambda_1^{\kappa},\cdots,\lambda_r^{\kappa},0,\cdots,0),\quad (P\Lambda P^{\tau})^{\kappa}=P\Lambda^{\kappa}P^{\tau}.$$

Let L_n be a $p\times p$ lower triangular matrix with positive diagonal elements. Then, decompositions

$$\Sigma_n^{-1}=L_n^{\tau}L_n,\quad \Sigma_n^{-1}=\Sigma_n^{-1/2}\Sigma_n^{-1/2},\quad \Sigma_n^{-1}=[D(n)^{-1/2}A(n)^{\tau}]^{\tau}[D(n)^{-1/2}A(n)^{\tau}] \qquad \ldots(2.1)$$

are LTD, SRD and JFD (cf. Section 1) of Σ_n^{-1} respectively. Based on these decomposition methods, we define $S_i(\cdot)$, $i=1,2,3$, by

$$S_1(X_n)=L_nX_n,\quad S_2(X_n)=\Sigma_n^{-1/2}X_n,\quad S_3(X_n)=D(n)^{-1/2}A(n)^{\tau}X_n \qquad \ldots(2.2)$$

and call them LTS, SRS and JFS respectively.

Remark 2.1. According to the above description, it is obvious that L_n and $\Sigma_n^{-1/2}$ are uniquely determined by Σ_n, but $A(n)$ is not, especially when Σ_n has eigenvalues of multiplicity bigger than one. This implies that the decomposition defined by the third equality in (2.1) is not uniquely determined. In the following, any such decomposition is called a JFD of Σ_n^{-1}. The corresponding $S_3(X_n)$ given in (2.2) is called a JFS of X_n. For data matrices X_n and Y_n, we write $S_3(X_n) \cong S_3(Y_n)$ if

$$\{S_3(X_n) : S_3(X_n) \text{ is a JFS of } X_n\} = \{S_3(Y_n) : S_3(Y_n) \text{ is a JFS of } Y_n\}.$$

2.1. Equivariance/Invariance and related. Subsections 2.1 and 2.2 study the equivariance/invariance properties of sphering under affine, orthogonal and lower triangular transformations, and applications to PP. A sphering S is called an orthogonal equivariant (invariant) sphering if $S(AX) = AS(X) \quad (S(AX) = S(X))$ for all X and orthogonal A. Similarly, we define the other kinds of equivariant/invariant spherings. In Subsections 2.1 and 2.2, the sample size n is fixed and the population dispersion matrix Σ is not involved, so we omit the subscripts n in X_n, Σ_n and B_n. That is, $\Sigma = \hat{\Sigma}(X)$ and $\Sigma^{-1} = B^\tau B$ is a decomposition of Σ^{-1}.

THEOREM 2.1. *Let $S_i(\cdot)$, $i = 1, 2, 3$, be defined in (2.2). We have:*

(i) For any orthogonal P and data matrix X, $S_2(PX) = PS_2(X)$, $S_3(PX) \cong S_3(X)$.

(ii) Let X be fixed with $rank(X) = p$. Then, (a) for any orthogonal P, $S_1(PX) = S_1(X)$ iff $P = I$; (b) $S_1(PX) = S_1(X)$ for all orthogonal P iff $\Sigma = \lambda I$ with some $\lambda > 0$; (c) for any nonsingular C, $S_1(CX) = S_1(X)$ iff C is lower triangular; (d) there exists a lower triangular matrix L with all diagonal elements equal to one such that $S_2(LX) \neq S_2(X)$; (e) for each JFS of X, $S_3(X)$, there exists a lower triangular matrix K with all diagonal elements equal to one such that $S_3(KX) \neq S_3(X)$ for any JFS of KX, $S_3(KX)$.

The proof of this theorem relies on the following two lemmas, which will be proved in Section 4.

LEMMA 2.1. *Let L be a $p \times p$ lower triangular matrix and A be a $p \times p$ matrix. Assume that L and A are both nonsingular. Then*

(i) LA is lower triangular iff A is lower triangular.

(ii) PLP^τ is lower triangular for every orthogonal P iff $L = \lambda I$ with some constant λ.

LEMMA 2.2. *For any $p \times p$ orthogonal $A, (p \geq 2)$ there exists a lower triangular L with all diagonal elements equal to one such that $A^\tau L^\tau LA$ is not diagonal.*

PROOF OF THEOREM 2.1. Conclusion (i) follows directly from the definitions of $S_2(\cdot)$ and $S_3(\cdot)$.

Note that $rank(X) = p$ implies that for any $p \times p$ matrices A and B

$$AX = BX \text{ iff } A = B. \qquad \ldots(2.3)$$

Thus, statements (a) and (b) in (ii) can be obtained by (2.3) and Lemma 2.1; and the statement (c) in (ii) follows immediately from the definition of $S_1(\cdot)$ and (2.3).

Now turn to statements (d) and (e) in (ii).

Denote $E(i,j) = (e_{kl})_{p\times p}$ with all $e_{kl} = 0$ except $e_{ij} = 1$. Using (2.3) it is easy to show that $L = I_p + E(i,j)$ $(i > j)$ is such that $S_2(LX) \neq S_2(X)$.

Let $S_3(X) = D^{-1/2}A^\tau X$ be a JFS of X. By Lemma 2.2, there is a lower triangular K with all diagonal elements being equal to one such that $A^\tau K^\tau KA$ is not diagonal. We show that $S_3(KX) \neq S_3(X)$ by contradiction. Suppose there is a $S_3(KX) = D_1^{-1/2}A_1^\tau KX$ such that $S_3(KX) = S_3(X)$. Then, by (2.3), $D_1^{-1/2}A_1^\tau K = D^{-1/2}A^\tau$. This gives $K^\tau K = AD^{-1}D_1A^\tau$. Thus $A^\tau K^\tau KA = D^{-1}D_1$ is diagonal, which results in a contradiction.

The proof is finished.

Remark 2.2. From Theorem 2.1, we see that $S_2(\cdot)$ is orthogonal equivariant and $S_3(\cdot)$ is orthogonal invariant; $S_1(\cdot)$ is neither equivariant nor invariant under orthogonal transformations; under lower triangular transformation, $S_1(\cdot)$ is invariant, whereas $S_2(\cdot)$ and $S_3(\cdot)$ are not. Thus, none of the commonly used three sphering procedures is affine invariant.

Remark 2.3. Let X be fixed. By a result in Rao (1973, Complements and Problem 3.2), we know that for any sphering transformation S and S^*, there exists a unique orthogonal $U = U(S, S^*)$ such that $S^*(X) = US(X)$. That is, the difference between two sphered data sets given by any two sphering transformations respectively (remember: $\hat{\Sigma}(\cdot)$ is fixed) is only a rotation, or a choice of axes.

2.2. When applied to PP. In PP analysis, an objective function called projection index, is previously chosen according to the purpose with which the data set is analyzed. Maximizing the projection index numerically based on the data set, the optimal directions and the corresponding optimal projections are obtained. Analyzing the projected data, structure or other aspects of the original data set are supposed to be revealed. For a comprehensive discussion of PP, see Huber (1985), Jones and Sibson (1987), and Li and Cheng (1993).

When structure or information other than contained in location/dispersion are of interest, one usually needs an affine invariant index. If an appropriate affine invariant index is not available, then sphering the data set before maximizing the index is a way to extract location/dispersion effects. In this subsection we compare these two procedures.

To simplify notation and give a clear insight, we limit ourselves to one-dimensional PP. As before, $X = (x_1, \cdots, x_n)$ is a data matrix of $p-$dimensional observations. Let Q be an index, i.e., a map from R^n to R. Denote $Q_M(X) = \max_{|a|=1} Q(a^\tau X)$, $a_M \hat{=} a_M(X)$ with $|a_M| = 1$ such that $Q(a_M^\tau X) = Q_M(X)$, where $|a|$ is the Euclidean norm of vector a. $a_M = a_M(X)$ is called the optimal direction of X. The corresponding (optimal) projection is $a_M^\tau X$. Note that the optimal direction may not be unique. In this case $a_M(X)$ can be any maxima. An index Q is said to be affine invariant if for any $a, b, y_1, \cdots, y_n \in R$ with $a \neq 0$,

$$Q(ay_1 + b, \cdots, ay_n + b) = Q(y_1, \cdots, y_n).$$

PROPOSITION 2.1. *If Q is affine invariant, then for any nonsingular $p \times p$ matrix C and data matrix X*

$$Q_M(CX) = Q_M(X), \quad a_M(CX)^\tau CX = \frac{a_M(X)^\tau X}{|(C^\tau)^{-1} a_M(X)|}.$$

The proof of this proposition is very simple and is omitted.

This proposition shows that the maximum value $Q_M(\cdot)$ of an affine invariant index is affine invariant, and up to a factor of positive constant, the optimal projections remain the same under affine transformations.

Now, let Q be any projection index and $S(\cdot)$ a sphering transformation defined by $S(X) = BX$. Then, PP-after-sphering yields that

$$Q_M(X, S) = \max_{|a|=1} Q(a^\tau S(X)),$$

$$a_M = a_M(X, S) \quad \text{with} \quad |a_M| = 1 \quad \text{satisfying} \quad Q(a_M^\tau S(X)) = Q_M(X, S).$$

The optimal projection now becomes $a_M^\tau S(X) = (B^\tau a_M)^\tau X$.

THEOREM 2.2. *In a PP-after-sphering procedure with index Q, we have:*

(i) For any X and $p \times p$ nonsingular C, $Q_M(CX, S) = Q_M(X, S)$; and for any two sphering procedures S and S^*, $Q_M(X, S) = Q_M(X, S^*)$.

(ii) For any data matrix X, nonsingular C and sphering procedures S and S^*,

$$a_M(CX, S)^\tau S(CX) = a_M(X, S)^\tau S(X) = a_M(X, S^*)^\tau S^*(X).$$

The proof of this theorem is easy.

This theorem shows that $Q_M(\cdot, S)$ is affine invariant. One advantage we take from it is that if properties of $Q_M(X, S)$ are studied, then we may choose a sphering S^* such that properties of $Q_M(X, S^*)$ is easier to derive. For example, the sphering matrix of $S_3(\cdot)$ does not have good asymptotic properties, but that of $S_2(\cdot)$ has, as shown in the next subsection. Substituting $S_2(\cdot)$ for $S_3(\cdot)$, we can give a proof for the limit distribution of the maximum value of Friedman PP index in which S_3 was adopted (see Friedman (1987), Sun (1989), Section 5.2, and Zhang (1995)).

Comparing this theorem with Proposition 2.1, we can see that the maximum values of projection indices have the same invariant property in both PP with affine equivariant index and PP-after-sphering, and that the optimal projections, which is what a PP procedure actually searches, are always affine invariant in both kinds of PP. Moreover, all PP-after-sphering procedures based on the same index give the same optimal projections. In this sense, all sphering transformations are equivalent when they are used in PP for searching optimal projections. However, when the asymptotic properties of the PP statistics Q_M and a_M are to be studied, different spherings may not perform in the same way. This is because their asymptotic properties are quite different as shown in the next subsection.

2.3 Asymptotic properties of sphering matrices. As stated in Section 1, sphering is applied as a useful pretreatment for data analyses. Then, formal statistical procedures are implemented on the sphered data sets. Thus, to study the performance of the whole statistical analysis, asymptotic properties of sphering are sometimes required. For example, in order to determine the P-values of a PP-after-sphering procedure, we need asymptotic distributions of $Q_M(X_n, S)$ and $a_M(X_n, S)$ (cf. Sun (1989), Zhang (1993)), which certainly depend on asymptotic properties of the sphering matrix adopted in sphering S. The commonly applied three type of sphering matrices based on LTD, SRD and JFD respectively are presented in (2.1). Their asymptotic properties are investigated in this subsection.

Let $DIAG(A_1, \cdots, A_r)$ denote a diagonal block matrix, and $(A_{ij})_{1\le i,j\le r}$ denote a partitioned matrix, i.e.,

$$DIAG(A_1, \cdots, A_r) = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_r \end{pmatrix}, \qquad \ldots (2.4)$$

$$(A_{ij})_{1\le i,j\le r} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ A_{r1} & A_{r2} & \cdots & A_{rr} \end{pmatrix}. \qquad \ldots (2.5)$$

In this subsection we assume that the population dispersion matrix Σ has a spectral decomposition

$$\Sigma = ADA^\tau, \quad D = DIAG(\lambda_1 I_{q_1}, \cdots, \lambda_r I_{q_r}), \quad \lambda_1 > \cdots > \lambda_r > 0, \quad \sum_{i=1}^{r} q_i = p \qquad \ldots (2.6)$$

and the estimator Σ_n of Σ based on data matrix X_n has a spectral decomposition

$$\begin{aligned} \Sigma_n &= A(n)D(n)A(n)^\tau, \quad D(n) = diag(d(n,1), \cdots, d(n,p)), \\ & \quad d(n,1) \ge \cdots \ge d(n,p) \ge 0, \end{aligned} \qquad \ldots (2.7)$$

where A and $A(n)$ are $p \times p$ orthogonal matrices, and q_i $(i = 1, \cdots, r)$ are positive integers. Also, "$\xrightarrow{P_r}$" means convergence in probability, "$\xrightarrow{d}$" means convergence in distribution, and "$\xrightarrow{a.s.}$" means convergence almost surely.

Theorem 2.3. *Assume that* $\Sigma_n \stackrel{P_r(a.s.)}{\longrightarrow} \Sigma$. *Let* L_n *and* $\Sigma_n^{-1/2}$ *be defined as in (2.1). Then*

(i) $L_n \stackrel{P_r(a.s.)}{\longrightarrow} L$, $\Sigma_n^{-1/2} \stackrel{P_r(a.s.)}{\longrightarrow} \Sigma^{-1/2}$. *That is,* L_n *and* $\Sigma_n^{-1/2}$ *are both consistent;*

(ii) $D(n)^{-1/2}A(n)^\tau$ *is not necessarily consistent. For example, if* Σ_n *is the sample covariance matrix of* $N(\mu,\Sigma)$, *then* $D(n)^{-1/2}A(n)^\tau$ *does not converge to* $D^{-1/2}A$ *in probability.*

Proof. (i) The consistency of L_n follows immediately from L_n being a continuous function of Σ_n. Partition $A(n)$ and A as

$$A(n) = (A_1(n), \cdots, A_r(n)), \quad A = (A_1, \cdots, A_r)$$

where $A_i(n)$ and A_i are $p \times q_i$ matrices, $i = 1, \cdots, r$. Denote $s_0 = 0$, $s_k = \sum_{i=1}^k q_i$. Thus,

$$\begin{aligned}\Sigma_n^{-1/2} - \Sigma^{-1/2} \quad &= \textstyle\sum_{k=1}^r \sum_{s_{k-1}<i\le s_k} (d(n,i)^{-1/2} - \lambda_k^{-1/2}) A_i(n)A_i(n)^\tau \\ &\quad + \textstyle\sum_{n=1}^r \lambda_k^{-1/2} \Big[\sum_{s_{k-1}<i\le s_k} A_i(n)A_i(n)^\tau - \sum_{s_{k-1}<i\le s_k} A_iA_i^\tau\Big].\end{aligned}$$

Because $d(n,i)$, $i = 1, \cdots, p$, and $\sum_{s_{k-1}<i\le s_k} A_i(n)A_i^\tau(n)$, $k = 1, \cdots, r$ are all continuous functions of Σ_n (see Kato, 1982, p.124), we conclude that $\Sigma_n^{-1/2} \stackrel{P_r(a.s.)}{\longrightarrow} \Sigma^{-1/2}$.

(ii) Firstly, $A(n)$ is not uniquely determined. One can choose a sequence of $A(n)$ such that $A(n)$ does not converge at all. In the case that Σ_n is the sample covariance matrix of $N(\mu,\Sigma)$, even if $A(n)$ is restricted by certain conditions so that it is uniquely determined, $A(n)$ converges, however, to a random matrix, rather than the constant matrix A (c.f. Anderson, 1963). Therefore, by Lemma 2.3 below, $D(n)^{-1/2}A(n)$ is not consistent.

Lemma 2.3. *If* $\Sigma_n \stackrel{P_r(a.s.)}{\longrightarrow} \Sigma$, *then* $D(n)^{-1/2}A(n)^\tau \stackrel{P_r(a.s.)}{\longrightarrow} D^{-1/2}A^\tau$ *implies* $A(n) \stackrel{P_r(a.s.)}{\longrightarrow} A$.

This lemma will be proved in Section 4.

Theorem 2.4. *Let* $U = (U_{ij})_{1\le i,j\le r}$ *be a* $p \times p$ *symmetric random matrix such that* U_{ij} *is a* $q_i \times q_j$ *matrix and each* U_{ii} *almost surely has distinct eigenvalues* $h(s_{i-1}+1) > \cdots > h(s_i)$, $i = 1, \cdots, r$, *where* $s_0 = 0$, $s_k = \sum_{i=1}^k q_i$. *Assume that* $c_n \longrightarrow \infty$ *and* $c_n(\Sigma_n - \Sigma) \stackrel{d}{\longrightarrow} V = AUA^\tau$. *Then, for any* $l \neq 0$, $c_n(\Sigma_n^l - \Sigma^l) \stackrel{d}{\longrightarrow} V^{(l)} = AU^{(l)}A^\tau$, *where* $U^{(l)} = (U_{ij}^{(l)})_{1\le i,j\le r}$, $U_{ii}^{(l)} = l\lambda_i^{l-1}U_{ii}$, $U_{ij}^{(l)} = \frac{\lambda_i^l - \lambda_j^l}{\lambda_i - \lambda_j} U_{ij}$. *Especially,* $c_n(\Sigma_n^{-1/2} - \Sigma^{-1/2}) \stackrel{d}{\longrightarrow} V^{(-1/2)}$.

To prove this theorem we need the following lemma.

Lemma 2.4. *Assume that (i)* D *is given in* (2.6)*; (ii)* $W = (W_{ij})_{1\le i,j\le r}$ *is a* $p \times p$ *(non-random) symmetric matrix,* W_{ij} *is* $q_i \times q_j$ *matrix,* $\sum_{i=1}^r q_i = p$, *each*

W_{ii} has distinct eigenvalues $h(s_{i-1}+1) > \cdots > h(s_i)$, with $s_0 = 0$, $s_k = \sum_1^r q_i$; (iii) $c_n \longrightarrow \infty$, non-random $S(n)$ is a $p \times p$ nonnegative definite matrix, $n = 1, 2, \cdots$, and $c_n(S(n) - D) \longrightarrow W$. Then, for any $l \neq 0$,

$$c_n(S(n)^l - D^l) \longrightarrow W^{(l)} = (W_{ij}^{(l)})_{1 \le i,j \le r}$$

where

$$W_{ii}^{(l)} = l\lambda_i^{l-1} W_{ii}, \quad W_{ij}^{(l)} = \frac{\lambda_i^l - \lambda_j^l}{\lambda_i - \lambda_j} W_{ij}, \quad i \neq j. \qquad \ldots (2.8)$$

Proof. See Section 4.

Proof of Theorem 2.4. Note that $c_n(\Sigma_n - \Sigma) = A[c_n(A^\tau \Sigma_n A - D)]A^\tau$ and $c_n(A^\tau \Sigma_n A - D) \xrightarrow{d} U$, We need only to prove

$$c_n(A^\tau \Sigma_n^l A - D^l) \xrightarrow{d} U^{(l)}. \qquad \ldots (2.9)$$

By Representation Theorem (see Pollard (1984), p. 71), there exist random matrices W and S_n, $n = 1, 2, \cdots$, such that

$$S_n \stackrel{d}{=} A^\tau \Sigma_n A, \quad W \stackrel{d}{=} U, \quad c_n(S_n - D) \xrightarrow{a.s.} W \qquad \ldots (2.10)$$

where "$\stackrel{d}{=}$" means both sides have the same distribution. Thus

$$c_n(A^\tau \Sigma_n^l A - D^l) \stackrel{d}{=} c_n(S_n^l - D^l). \qquad \ldots (2.11)$$

Partition $W = (W_{ij})_{1 \le i,j \le r}$ in the same way as $U = (U_{ij})_{1 \le i,j \le r}$. It follows that the eigenvalues of W_{ii}, denoted by $g(j)$, $s_{i-1} < j \le s_i$, $i = 1, \cdots, r$, satisfy $(g(s_{i-1}+1), \cdots, g(s_i)) \stackrel{d}{=} (h(s_{i-1}+1), \cdots, h(s_i))$, $i = 1, \cdots, r$, and therefore $g(s_{i-1}+1) > \cdots > g(s_i)$ *a.s.*. Applying Lemma 2.4 we obtain that

$$c_n(S_n^l - D^l) \xrightarrow{a.s.} W^{(l)} = (W_{ij}^{(l)})_{1 \le i,j \le r},$$

with W_{ij}, $1 \le i, j \le r$, given in (2.8). From (2.10) it is obvious that $W^{(l)} \stackrel{d}{=} U^{(l)}$. Then, by (2.11) we have (2.9). The proof of the theorem is completed.

Before giving the limit distribution of the sphering matrix L_n based on LTD, we state a few symbols and results needed below.

Let $A \otimes B$ be the Kronecker product of matrices A and B, $Vec(A) = (a_1^\tau, \cdots, a_t^\tau)^\tau$ the vec-function of $s \times t$ matrix $A = (a_1, \cdots, a_t)$, $e_j(s)$ a s-vector with a one in jth position and zeroes elsewhere, and $E(i,j)$ a $p \times p$ matrix with a one in (i,j) position and zeroes elsewhere. Denote $I_(p,p) = \sum_{i=1}^p \sum_{j=1}^p E(i,j) \otimes E(j,i)$. Suppose that A, B and C are, respectively, $s \times t$, $t \times u$, $u \times v$ matrices, and that D and E are both $p \times p$ matrices. The following three equalities (cf. Neudecker (1968), Magnus and Neudecker (1979)) are useful tools.

$$\begin{aligned} Vec(ABC) &= (C^\tau \otimes A)Vec(B), \quad I_{(p,p)}Vec(D) = Vec(D^\tau), \\ I_{(p,p)}(D \otimes E) &= (E \otimes D)I_{(p,p)}. \end{aligned} \qquad \ldots (2.12)$$

Also, denote $I_{p^2} + I_{(p,p)} = (d_1, \cdots, d_{p^2})^\tau$. It is easy to verify that

$$\begin{aligned} d_{(j-1)p+i} &= e_{(j-1)p+i}(p^2) + e_{(i-1)p+j}(p^2) = e_j(p) \otimes e_i(p) + e_i(p) \otimes e_j(p) \\ &= d_{(i-1)p+j}. \end{aligned} \qquad \ldots (2.13)$$

THEOREM 2.5. *Assume that* $c_n \longrightarrow \infty$ *and* $c_n(\Sigma_n - \Sigma) \xrightarrow{d} V$. *Let* L_n *and* L *be* $p \times p$ *lower triangular such that* $\Sigma_n^{-1} = L_n^\tau L_n$, $\Sigma^{-1} = L^\tau L$. *Denote* $L^{-1} = (m_{ij})_{p\times p}$; $\overline{v}_{ij} = 0$, $1 \le i < j \le p$, $\overline{v}_{ij} = v_{ij}$, $1 \le j \le i \le p$; *and* $\overline{V} = (\overline{v}_{ij})_{p\times p}$. *Then:*

(i) There exists a random matrix Z *such that* $c_n(L_n^{-1} - L^{-1}) \xrightarrow{d} Z$ *and* $c_n(L_n - L) \xrightarrow{d} -LZL$.

(ii) Z *is uniquely determined by the following recurrence formulas:*

$$\begin{aligned} z_{ts} &= 0,\ 1 \le t < s \le p; \quad z_{t1} = \frac{1}{m_{11}}(v_{t1} - \frac{m_{t1}}{2m_{11}}v_{11}),\ 1 \le t \le p; \\ z_{ts} &= \frac{1}{m_{ss}}(v_{ts} - \frac{m_{ts}}{2m_{ss}}v_{ss} + \frac{m_{ts}}{m_{ss}}\sum_{k=1}^{s-1} z_{sk}m_{sk} - \sum_{k=1}^{s-1}(z_{tk}m_{sk} + m_{tk}z_{sk})), \\ &\qquad 1 \le s \le t \le p. \end{aligned}$$

(iii) Set $G = (g_1, \cdots, g_{p^2})^\tau$ *with* $g_{(j-1)p+i} = e_j^\tau(p)L \otimes e_i(p)$, $\quad 1 \le i < j \le p$; $g_{(j-1)p+i} = d_{(j-1)p+i}$, $1 \le j \le i \le p$; *where* d_k $(1 \le k \le p^2)$ *are given in* (2.13). *We have* $Vec(Z) = (L \otimes I_p)G^{-1}Vec(\overline{V})$, $Vec(-LZL) = -(\Sigma^{-1} \otimes L)G^{-1}Vec(\overline{V})$.

PROOF OF THEOREM 2.5. First, assertion (i) follows directly from the continuous differentiability of L_n^{-1} as a function of Σ_n, and the argumentation by the standard $\delta-$method. Then, from assertion (i) it follows that $c_n(\Sigma_n - \Sigma) \xrightarrow{d} ZL^{-\tau} + L^{-1}Z^\tau$. This yields

$$ZL^{-\tau} + L^{-1}Z^\tau = V. \qquad \ldots (2.14)$$

Rewrite (2.14) in elements we obtain, with $i \wedge j = \min\{i, j\}$,

$$v_{ij} = \sum_{k=1}^{i\wedge j}(z_{ik}m_{jk} + m_{ik}z_{jk}).$$

This, together with Z being lower triangular, gives inductively assertion (ii). Finally, applying (2.12), (2.14) and $e_i^\tau(p)Ze_j(p) = 0$, $1 \le i < j \le p$, we obtain

$$\begin{aligned} Vec(V) &= (L^{-1} \otimes I_p)Vec(Z) + (I_p \otimes L^{-1})Vec(Z^\tau) \\ &= [L^{-1} \otimes I_p + (I_p \otimes L^{-1})I_{(p,p)}]Vec(Z) \\ &= (I_{p^2} + I_{(p,p)})(L^{-1} \otimes I_p)Vec(Z). \end{aligned} \qquad \ldots (2.15)$$

$$[(e_j^\tau(p)L) \otimes e_i^\tau(p)](L^{-1} \otimes I_p)Vec(Z) = 0, \quad 1 \le i < j \le p. \qquad \ldots (2.16)$$

Note the reappearance of the $[(j-1)p+i]$th element of $Vec(V)$ $(1 \le i < j \le p)$ and the corresponding rows of $I_{p^2} + I_{(p,p)}$, we replace them by zeroes in V and by $e_j^\tau(p)L \otimes e_i^\tau(p)$ $(1 \le i < j \le p)$ in $I_{p^2} + I_{(p,p)}$. Then (2.15) and (2.16) becomes

$$G(L^{-1} \otimes I_p)Vec(Z) = Vec(\overline{V}). \qquad \ldots (2.17)$$

It is easy to see that G is nonsingular. Thus, the assertion (iii) follows from (2.17), (2.12) and $(L^{-1} \otimes I_p)^{-1} = L \otimes I_p$.

REMARK 2.4. If we omit those zeroes in $Vec(\overline{V})$, then, the equality $Vec(Z) = (L \otimes I_p)G^{-1}Vec(\overline{V})$ can be simplified. Specifically, let A_1 be the $(p(p+1)/2) \times p^2$ matrix obtained by deleting the $[(j-1)p+i]$th $(1 \le i < j \le p)$ rows of $I_{p^2} + I(p,p)$; and B_1 the $p^2 \times (p(p-1)/2)$ matrix obtained by adding $[(j-1)p+i]$th column to $[(i-1)p+j]$th column $(1 \le i < j \le p)$, and then deleting the $[(j-1)p+i]$th $(1 \le i < j \le p)$ columns in $L^{-1} \otimes I_p$. Denote

$$\begin{aligned} Vecs(Z) &= (z_{11}, \cdots, z_{p1}; z_{22}, z_{32}, \cdots, z_{p2}; \cdots; z_{pp})^\tau, \\ Vecs(V) &= (v_{11}, \cdots, v_{p1}; v_{22}, v_{32}, \cdots, v_{p2}; \cdots; v_{pp})^\tau. \end{aligned}$$

Then, $Vecs(Z) = (A_1B_1)^{-1}Vecs(V)$.

3. Discussion

Centering and sphering is intended to remove location, scale and correlation structure and focus one's attention on other types of structures. All the centering and sphering procedures defined in this paper serve this purpose since centered and sphered data sets exclusively have zero location, unit scale, and zero correlation coefficients (it follows evidently from their identity dispersion matrices). The difference between any two sphering procedures lies in a rotation (cf. Remark 2.3), which certainly has no effects on location, scale, and correlations. Also, in PP-after-sphering procedures, all sphering transformations give the same optimal projections, from which analysts try to find structures in the data.

On the other hand, centering and sphering is only an initial stage of certain types of data analyses. Performance of further analysis may depend on the behaviour of the initial procedure. From this viewpoint, we recommend the SRS (square root sphering) and LTS (lower triangular sphering). SRS not only is always uniquely determined by the dispersion matrix (even in the case that the sample dispersion matrix is singular), but also has several good properties. It possesses orthogonal equivariance (Theorem 2.1), and consistency if the sample dispersion matrix is consistent (Theorem 2.3), and, most importantly, the asymptotic distribution of its sphering matrix can be easily computed from that of the sample dispersion matrix (Theorem 2.4). The LTS has also some advantages. A conspicuous one is ease of computation. It is unique if the sample dispersion matrix is positive definite and converges in distribution if the sample dispersion matrix does.

4. **Proof of Lemmas 2.1–2.4**

This section technically verifies the lemmas presented in Section 2.

PROOF OF LEMMA 2.1. Conclusion (i) and the necessary part of (ii) are easy, we now verify the sufficient part of (ii).

Suppose that PLP^τ is lower triangular for any orthogonal P. Denote $P = (P_1, \cdots, P_p)$. Then

$$P_i^\tau LP_j = 0, \quad \text{for} \quad j > i. \qquad \ldots (4.1)$$

Let $e_j = (0, \cdots, 1, \cdots, 0)^\tau$ be the $p-$vector with a one in the jth position and zeroes elsewhere, $1 \le j \le p$. Set $P = (e_p, e_{p-1}, \cdots, e_1)$. Then (4.1) yields that $l_{ij} = 0$ for $i < j$. Hence, $L = diag(l_{11}, \cdots, l_{pp})$.

To complete this lemma, we need only to show that

$$l_{11} = l_{22} = \cdots = l_{pp}. \qquad \ldots (4.2)$$

Suppose that there are two different diagonal elements in L. Without loss of generality, we assume that $l_{11} \neq l_{22}$. Set

$$A_1 = \begin{pmatrix} \sin\alpha & \cos\alpha \\ \cos\alpha & -\sin\alpha \end{pmatrix}, \quad A = DIAG(A_1, A_2)$$

where A_2 is a $(p-2)\times(p-2)$ orthogonal matrix. Thus, A is orthogonal for any α. Then by assumption ALA^τ is lower triangular. This yields that

$$\begin{aligned} 0 &= (\sin\alpha, \cos\alpha) diag(l_{11}, l_{22})(\cos\alpha, -\sin\alpha)^\tau \\ &= (l_{11} - l_{22})\sin\alpha\cos\alpha, \quad \text{for any} \quad \alpha. \end{aligned}$$

Hence $l_{11} = l_{22}$. This produces a contradiction. Therefore (4.2) holds.

PROOF OF LEMMA 2.2. Write $A = (A_1, \cdots, A_p) = (a_{ij})_{p\times p}$. Obviously, $A_1 \neq 0$, without loss of generality we assume $a_{11} \neq 0$. Consider the following two cases.

Case 1 : $a_{j1} = 0$ for all $j = 2, \cdots, p$. Denote $E(i,j) = (e_{kl})_{p\times p}$ with all $e_{kl} = 0$ except $e_{ij} = 1$. We claim that $L = I_p + cE(2,1)$ $(c \neq 0)$ is required. Actually, if $A^\tau L^\tau LA \hat{=} \Delta \hat{=} diag(\delta_1, \cdots, \delta_p)$ is diagonal, then, obviously, $\Delta > 0$, $L^\tau LA_1 = \delta_1 A_1$ and $L^\tau L = I_p + cE(2,1) + cE(1,2) + c^2E(1,1)$. Thus, it follows that $0 = \delta_1 a_{21} = ca_{11} + a_{21} = ca_{11}$. Hence $a_{11} = 0$, which contradicts with $a_{11} \neq 0$. Therefore $A^\tau L^\tau LA$ cannot be diagonal.

Case 2 : there is a j, $2 \le j \le p$ such that $a_{j1} \neq 0$. Without loss of generality, assume that $a_{21} \neq 0$. Then, for $c > 0$ large enough, the above L is still such that $A^\tau L^\tau LA$ is not diagonal. Otherwise, by the same argument we obtain that

$$(c^2+1)a_{11} + ca_{21} = \delta_1 a_{11}, \quad ca_{11} + a_{21} = \delta_1 a_{21}.$$

It follows that $c^2 + 1 + ca_{21}/a_{11} = ca_{11}/a_{21} + 1$, which can not hold for $c > 0$ large enough. The proof is completed.

PROOF OF LEMMA 2.3. Since eigenvalues of a matrix are continuous functions of the matrix, $\Sigma_n \xrightarrow{P_r} \Sigma$ implies that $D(n)^{-1/2} - D^{-1/2} = o_p(1)$. Thus

$$\begin{aligned} o_p(1) &= D(n)^{-1/2}A(n)^\tau - D^{-1/2}A^\tau \\ &= (D(n)^{-1/2} - D^{-1/2})A(n)^\tau + D^{-1/2}(A(n) - A)^\tau \\ &= o_p(1) + D^{-1/2}(A(n) - A)^\tau. \end{aligned}$$

This gives that $A(n) \xrightarrow{Pr} A$. The proof for almost sure convergence is the same.

PROOF OF LEMMA 2.4. We follow the notation introduced in (2.4) and (2.5). Let $d(n,1) \ge \cdots \ge d(n,p)$ be the eigenvalues of $S(n)$. Denote

$$\begin{aligned} H_k &= diag(h(s_{k-1}+1), \cdots, h(s_k)),\ k = 1, \cdots, r; \\ H &= DIAG(H_1, \cdots, H_r); \\ D_k(n) &= diag(d(n, s_{k-1}+1), \cdots, d(n, s_k)),\ k = 1, \cdots, r; \\ D(n) &= DIAG(D_1(n), \cdots, D_r(n)); \\ D &= DIAG(\lambda_1 I_{q_1}, \cdots, \lambda_r I_{q_r}) = (D_{ij})_{1 \le i,j \le r}, \end{aligned}$$

where D_{ij} is $q_i \times q_j$ matrix. Also, all (i, j) $(1 \le i, j \le r)$ blocks of matrices in this proof are $q_i \times q_j$ matrices.

It follows from Anderson (1989, Theorem 4.1) that

$$\lim c_n(D(n) - D) = H; \qquad \ldots (4.3)$$

and that there exist orthogonal matrices $B(n) = (B_{ij}(n))_{1 \le i,j \le r}$, $n = 1, 2, \cdots$, and $B = (B_{ij})_{1 \le i,j \le r}$ such that

$$\begin{aligned} &B_{ij} = 0, i \neq j, \quad BHB^\tau = DIAG(W_{11}, \cdots, W_{rr}), \\ &B(n)D(n)B(n)^\tau = S(n), \quad \lim B(n) = B; \end{aligned} \qquad \ldots (4.4)$$

$$\lim c_n B_{ij}(n) \quad \text{exists, denoted by} \quad M_{ij}, \quad \text{for} \quad i \neq j; \qquad \ldots (4.5)$$

$$\begin{aligned} &\lim c_n(B_{kk}(n)B_{kk}(n)^\tau - I_{q_k}) = 0; \\ &B_{ii}M_{ji}^\tau = -M_{ij}B_{jj}^\tau = W_{ij}/(\lambda_i - \lambda_j). \end{aligned} \qquad \ldots (4.6)$$

Note that

$$c_n(S(n)^l - D^l) = c_n B(n)(D(n)^l - D^l)B(n)^\tau + c_n[B(n)D^l B(n)^\tau - D^l].$$

From (4.3) it follows that

$$\lim c_n(d(n,i)^l - \lambda_k^l) = l\lambda_k^{l-1}h(i), \quad s_{k-1} < i \le s_k, k = 1, \cdots, r,$$

or, equivalently $\lim c_n(D(n)^l - D^l) = DIAG(l\lambda_1^{l-1}H_1, \cdots, l\lambda_r^{l-1}H_r)$. Thus,

$$\lim c_n B(n)(D(n)^l - D^l)B(n)^\tau = DIAG(W_{11}^{(l)}, \cdots, W_{rr}^{(l)}).$$

Denote $T(n) = B(n)D^lB(n)^\tau - D^l = (T_{ij}(n))_{1\le i,j\le r}$. Then, by (4.4)~(4.6),

$$\begin{aligned}\lim c_n T_{kk}(n) &= \lim c_n \lambda_k^l (B_{kk}(n)B_{kk}(n)^\tau - I_{q_k}) = 0, \quad k = 1, \cdots, r.\\ \lim c_n T_{ij}(n) &= \lim c_n \textstyle\sum_{k=1}^r \lambda_k^l B_{ik}(n)B_{jk}(n)^\tau = \lambda_i^l B_{ii}M_{ji}^\tau + \lambda_j^l M_{ij}B_{jj}^\tau\\ &= (\lambda_i^l - \lambda_j^l)W_{ij}/(\lambda_i - \lambda_j), i \ne j.\end{aligned}$$

Therefore $\lim c_n(S(n)^l - D^l) = W^{(l)}$.

This proof is completed.

Acknowledgements. The authors would like to thank the referee for his helpful comments. This work is supported by National Natural Science Foundation of China and by Postdoctoral Foundation of China.

References

ANDERSON, T.W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34**, 122-148.

———— (1989). The asymptotic distribution of characteristic roots and vectors in multivariate components of variance. In *Contributions to Probability and Statistics, Essays in Honor of Ingram Olkin*, L.J. Gleser, M.D. Perlman, S.T. Press, and A.R. Sampson eds, p.177-196.

FRIEDMAN, J.H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82**, 249-266.

HUBER, P.J. (1981). Projection pursuit. *Research Report PJH-6*, Department of Statistics, Harvard University.

———— (1985). Projection pursuit (with discussions). *Ann. Statist.* **13**, 435-475.

JONES, M.C. AND SIBSON, R. (1983). What is projection pursuit? (with discussions). *J. R. Statist. Soc. A* **150**, 1-36.

KATO, T. (1982). *A Short Introduction to Perturbatory Theory for Linear Operators.* Springer-Verlag, New York.

LI, G. AND CHENG, P. (1993). Some recent developments in projection pursuit in China. *Statistica Sinica* **3**, 35-52.

MAGNUS, J.R. AND NEUDECKER, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.* **7**, 381-394.

NEUDECKER, H. (1968). The Kronecker matrix product and some of its applications. *Statistica Neerlandica* **23**, 69-82.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications (2nd edition).* John Wiley & Sons, New York.

SUN, J. (1989). P-values in projection pursuit. *Technical Report No.104*, Department of Statistics, Stanford University.

TUKEY, P.A. AND TUKEY, J.W. (1981). Graphical display of data in three and higher dimensions. In *Interpreting Multivariate Data,* V. Barnett, ed., Wiley, New York.

ZHANG, J. (1993). The mode index for PP clustering and the asymptotics on estimation of its optimum direction. *Acta Mathematicae Applicatae Sinica*, **16**, 171-184.

−−−− (1995). Data sphering: some properties and applications. *Technical Report*, Institute of Systems Science, Academia Sinica.

GUOYING LI AND JIAN ZHANG
INSTITUTE OF SYSTEMS SCIENCE
ACADEMIA SINICA
BEIJING 100080
CHINA.
email : jzhang@stat.ucl.ac.be