Published in final edited form as:
Comput Stat Data Anal. 2008 January 20; 52(5): 2292–2310. doi:10.1016/j.csda.2007.09.012.

An overview of statistical decomposition techniques applied to complex systems

Yalcin Tuncer,
Professor Emeritus, Middle East Technical University and Ankara University, Ankara, Turkey

Murat M. Tanik*, and
Department of Electrical and Computer Engineering, U.A.B., Birmingham, Alabama 35294-4461, USA

David B. Allison
Department of Biostatistics, School of Public Health, U.A.B., Birmingham, Alabama

Abstract

The current state of the art in applied decomposition techniques is summarized within a comparative uniform framework. These techniques are classified by the parametric or information theoretic approaches they adopt. An underlying structural model common to all parametric approaches is outlined. The nature and premises of a typical information theoretic approach are stressed. Some possible application patterns for an information theoretic approach are illustrated. Composition is distinguished from decomposition by pointing out that the former is not a simple reversal of the latter. From the standpoint of application to complex systems, a general evaluation is provided.

Keywords

Bipartite network; Blind source separation; Complexity; Composition; Entropy; Independent component analysis; Information; Information transfer; Integration; Mutual information; Negentropy; Network component analysis; Principal component analysis; Singular value decomposition

Introduction

Decomposition is a process of breaking up into constituent elements. In mathematical analysis, it means factorization and/or finding summands of a real number or a matrix. In systems science, decomposition consists of finding an optimal partition of a system in terms of its subsystems. Decompositions in real-life applications are motivated by a need to obtain a much simpler body of constituents that can best represent a given system of unmanageable size and/or complex structure. Complexity is a lack of information about a system, and unmanageable size means high dimensionality (Donoho, 2000; Fan and Li, 2006). Optimality of decomposition is evaluated by means of some adopted criteria, such as a dispersion measure of observables (for instance, higher eigenvalues or singular values of covariance matrix) or

*Corresponding author. Phone: 205 934 8442, Fax: 205 975 3337, e-mail: mtanik@uab.edu.

their conformity with prior knowledge on network structure or an entropic measure (information content) involved, etc.

Here we aim to discuss the current state of the art in decomposition within a uniform framework of approach. We do not intend this to be a full-fledged survey. For illustrations and examples on some known techniques, readers will be referred elsewhere. Because decomposition has a vast spectrum of application areas, finding a common framework of interest to all is challenging. The case is the same with determining the type of audience to be addressed. We have chosen a general statistical and information theoretic framework without dwelling on specifics of a certain area, for example, technicalities of statistical inference issues relating to the models discussed. Apart from taking care to introduce a conceptual framework for decomposition on the basis of partitions and except for the two illustrations provided on applications of the last technique and our efforts to find a common framework for exposition, we do not claim originality:

A close study of the relevant literature has led us to conclude that applications of decomposition and relevant techniques fit into two categories of approaches: The first category handles the problem in terms of a structural formal model representing the real-world phenomenon to be considered. The second category uses information theoretic treatment of systems without specifying such a structure. The first four well-known techniques to be discussed can be placed within the first category and are also used for dimension reduction: principal component analysis (PCA), singular value decomposition (SVD), independent component analysis (ICA), and network component analysis (NCA). The final technique, which may be called decomposition with information transfer function (ITF), falls in the second category. Although it is one of the earliest approaches found in literature, ITF is little known and the least explored. In the exposition below, both random vectors and matrices are shown in bold uppercase letters whereas bold italic uppercase letters are used for random vectors only. The superscript t stands for transposition.

A common framework of structure for the first category of techniques can be set up in terms of a linear model such as

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n1} & \dots & a_{nm} \end{bmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix} \quad (1)$$

In information theory, for instance, this model represents a simple formal communication channel without encoding and decoding (Ash, 1990, Chapter I). The model (1) will shortly be expressed as

$$\boldsymbol{X} = \mathbf{A}\boldsymbol{Z} + \boldsymbol{E}$$

The column vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^t$ of the model (1) has n observable random variables $X_1, X_2, \dots, X_n$ such as phenotypes in biology. The coordinates of the column vector $\boldsymbol{Z}$ stand for m non-observable random factors $Z_1, Z_2, \dots, Z_m$, with $n \leq m$, where m denotes the number of arrays into which these latent factors (say, genotypes in the case of the given biological example) can be placed. The $(n \times 1)$ vector $\boldsymbol{E}$ contains n unobserved residuals $E_1, E_2, \dots, E_n$, which are sometimes called noise factors. When $E_1, E_2, \dots, E_n$ are missing in the model, i.e., when $\boldsymbol{E} = \boldsymbol{0}$, (1) becomes a noise-free or noiseless model. Setting $\boldsymbol{Y} = \mathbf{A}\boldsymbol{Z}$ and having hence $\boldsymbol{X}$

= $\boldsymbol{Y} + \boldsymbol{E}$ for (1), one easily obtains an initial decomposition $\{\boldsymbol{Y}, \boldsymbol{E}\}$ of $\boldsymbol{X}$ underlying the model (1). The assumption behind the decomposition $\{\boldsymbol{Y}, \boldsymbol{E}\}$ is that these two vectors or factors have additive (independent or at least orthogonal) effects on $\boldsymbol{X}$. $\boldsymbol{Y} = \mathbf{A}\boldsymbol{Z}$ is a linear approximation to any differentiable non-linear function (correspondence) $\boldsymbol{Y} = \Psi(\boldsymbol{Z})$ at a certain point of $\boldsymbol{Z}$. The general setup (1) outlined above is different from the classical multivariate regression model. In the multivariate regression model, the matrix **A** is observable and the vector $\boldsymbol{Z}$ is unobservable. Furthermore, under the regression model, the noise vector $\boldsymbol{E}$ cannot obviously be neglected.

The vectors $\boldsymbol{Z}$ and $\boldsymbol{X}$ are related to each other though an unknown mixing matrix **A,** the rows of which correspond to the coordinates of $\boldsymbol{X}$, i.e., phenotypes in biology, and the columns correspond to the elements of $\boldsymbol{Z}$. In physical terms, matrix **A** stands for a system or a network (i.e., gates) through which, for example, genotypes or similar latent factors are connected to phenotypes or observable outputs. Often, these matrices give some idea about the underlying structure of a system. Such matrices are sometimes given special names like design structure matrices, dependency structure matrices, problem solving matrices, or design precedence matrices (Browning, 2001). In fact, a lower triangular matrix **A** corresponds to a hierarchical structure and diagonal or block diagonal matrices indicate independence or block-wise independence of components. Generally, matrix cells contain real or complex numbers. In a bipartite network system (e.g., a network between two different groups such as the phenotypes and genotypes in our biology example), however, they are composed of non-negative integers only. Irrespective of their numerical nature, estimation of or information on such matrices becomes of utmost importance for the study of complex systems.

The main statistical problem tackled by the structural approach in (1) is thus to obtain a statistically significant estimate of the matrix **A** and hence the vector $\boldsymbol{Z}$ by using the information contained in the observable $\boldsymbol{X}$. In the statistical identification and estimation of matrix **A,** vector $\boldsymbol{X}$ is assumed generally to have a statistical distribution with non-spherical contours on its domain (e.g., if the variance-covariance matrix of a multivariate normal distribution has distinct eigenvalues, the corresponding probability density function traces ellipsoids in its domain whereas a covariance matrix with identical eigenvalues produces spheroids in its domain). PCA and SVD are basically designed for distributions with non-spherical contours; ICA works well with non-normal distributions. Furthermore, normal distributions have maximum entropy among other distributions with an identical mean and variance and therefore are inappropriate for entropic analyses that seek minimum entropy.

The first four techniques to be discussed below aim at obtaining matrix **A** from parameters of relevant distributions or at estimating it in terms of some observational phenomena. For the purpose of estimation, some finite number of observations like $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$, $n < N$, on $\boldsymbol{X}$ are obtained. The condition $n < N$ is obviously required for non-singularity. Hence, ignoring the corresponding residual vectors $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_N$ for the corresponding vectors $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ for the time being, we have the observational phenomenon

$$\begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1N} \\ X_{21} & X_{22} & \ldots & X_{2N} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nN} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n1} & \ldots & a_{nm} \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} & \ldots & Z_{1N} \\ Z_{21} & Z_{22} & \ldots & Z_{2N} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ Z_{m1} & Z_{m2} & \ldots & Z_{mN} \end{bmatrix} \quad (2)$$

In matrix notation, (2) can be re-expressed as

$$\mathbf{X}=\mathbf{AZ} \tag{3}$$

The matrix **X** corresponds to a set of *N* observations on *n* genes, for instance. The objective for the first four techniques is to obtain an estimate of the unknown structure matrix **A** by means of the observable **X** and hence to obtain information on **Z** with the model (2).

As mentioned earlier, ITF does not involve such a structure as (1) above. Instead, an input process like the latent ***Z*** of the above discussion is assumed to be transformed to an output process like the observable ***Y*** (or ***X***) above with the understanding that ***Z*** and ***Y*** (or ***X***) may not necessarily obey vector space algebra or any other space algebra with certain topological properties. Both the input and the output of transformation are assumed to be stochastic. We denote a system with $S=\{X_1, X_2,\ldots,X_n\}$, meaning that $X_1, X_2,\ldots,X_n$ are just components of the system; *S* does not represent a functional notation. As it is the case with a structural model as in (1), $X_1, X_2,\ldots,X_n$ usually correspond to output. Because we are interested in the nature of a system, this output feature of $X_1, X_2,\ldots,X_n$ is often overlooked. Further discussion on systems can be found in the treatment of ITF below. The objective of this final technique is thus to find a partition $\boldsymbol{p}\{S_1, S_2, \ldots,S_q\}$, $q \leq n$ (e.g., a set of exhaustive disjoint subsets $S_1, S_2, \ldots,S_q$), that provides the same information as *S*.

Principal Component Analysis

Although the underlying mathematical tool of PCA is not new, its application to statistical problems and its subsequent independent development are attributed to Pearson (1901) and Hotelling (1933). This mathematical tool is known as spectral decomposition of nonsingular (positive- or negative-definite) symmetrical square matrices and is sometimes referred as the Hoteling or Karhunen-Loève transform. Because a variance-covariance of a random vector satisfies these square-symmetry and positive-definiteness properties, spectral decomposition can also apply to the variance-covariance matrix $V(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{\varepsilon}(\boldsymbol{X}-\boldsymbol{\varepsilon}(\boldsymbol{X}))(\boldsymbol{X}-\boldsymbol{\varepsilon}(\boldsymbol{X}))^{t}$ of the observable vector ***X*** in model (1) or its estimate $\hat{\boldsymbol{\Sigma}}$ obtained from observation matrix **X** in (2) in the usual way:

$$\widehat{\sum}=\frac{1}{N}(\mathbf{X}-\frac{1}{N}\mathbf{X}\,l)(\mathbf{X}-\frac{1}{N}\mathbf{X}\,l)^{t}$$

where ***l*** is an $(N\times 1)$ sum vector, i.e., $\boldsymbol{l}=(1,1,\ldots,1)^{t}$, and as pointed out above, we have the condition $n < N$ for $\hat{\boldsymbol{\Sigma}}$ to be positive definite (Seber, 1984, p. 59). We shall henceforth deal with such positive-definite cases and shall refer to them as standard PCA. The symbol ^ above any character that denotes a parameter stands (conventionally) for the corresponding estimate.

Spectral decomposition of $\boldsymbol{\Sigma}$ or $\hat{\boldsymbol{\Sigma}}$ is then the factoring-out of these matrices such as

$$\sum=\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{t}\text{ and }\widehat{\sum}=\widehat{\mathbf{Q}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{Q}}^{t}$$

where **Q** is an $n\times n$ matrix with orthonormal columns $\boldsymbol{Q}_1, \boldsymbol{Q}_2,\ldots, \boldsymbol{Q}_n$ that are basically the eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix with positive diagonal elements $\lambda_1, \lambda_2,\ldots, \lambda_n$ consisting of the corresponding eigenvalues. We also have similar matrices $\hat{\mathbf{Q}}$ and $\hat{\boldsymbol{\Lambda}}$ for $\hat{\boldsymbol{\Sigma}}$. In view of the orthonormality of $\boldsymbol{Q}_1, \boldsymbol{Q}_2,\ldots, \boldsymbol{Q}_n$, the following obvious results are obtained:

$$\mathbf{Q}\mathbf{Q}^t=\mathbf{Q}^t\mathbf{Q}=\mathbf{I}=\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{4}$$

The same results will obviously hold for its sample counterpart $\hat{\mathbf{Q}}$. The spectral decomposition of $\mathbf{\Sigma}$ can be re-expressed in the better known way as

$$\sum=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^t=\sum_{k=1}^{n}\lambda_k\, Q_k Q_k{}^t=\ddagger\!\!\!\!\sum_{k=1}^{n}\sum_{\mathrm{k}} \tag{5}$$

The last equality shows that the variance-covariance matrix can also be decomposed as the sum of n matrices $\mathbf{\Sigma}_k=\lambda_k\, \boldsymbol{Q}_k\boldsymbol{Q}_k^t$, ($k$=1,2,…,n), which are orthogonal to each other so that $\mathbf{\Sigma}_i\mathbf{\Sigma}_j=\mathbf{0}$ for $i \neq j$. Obviously, the matrices $\mathbf{Q}_k = \boldsymbol{Q}_k\, \boldsymbol{Q}^t{}_k$ are idempotent (orthogonal projection) matrices because

$$\mathbf{Q}\mathbf{Q}^t=\ddagger\!\!\!\!\sum_{k=1}^{n}\mathbf{Q}_k=\ddagger\!\!\!\!\sum_{k=1}^{n}\boldsymbol{Q}_k\, \boldsymbol{Q}^t{}_k=\mathbf{I} \tag{6}$$

The spectral decomposition $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^t$ introduces further computational conveniences for inversion and squaring the variance-covariance matrix $\mathbf{\Sigma}$, i.e.,

$$\sum{}^2=\ddagger\!\!\!\!\sum_{k=1}^{n}(\lambda_k)^2\, \boldsymbol{Q}_k\boldsymbol{Q}_k{}^t \text{ and } \sum{}^{-1}=\ddagger\!\!\!\!\sum_{k=1}^{n}(\lambda_k)^{-1}\, \boldsymbol{Q}_k\boldsymbol{Q}_k{}^t$$

Parallel statements will be valid for the sample counterpart $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ as well. (For other aspects of spectral decomposition of variance-covariance matrices, see Jolliffe, 2002).

Such decomposition has some implications in statistical applications for the case that $n = m$. In fact, when $\mathbf{\Sigma}$ or $\hat{\mathbf{\Sigma}}$ is known, and in consequence when $\mathbf{Q}$ or $\hat{\mathbf{Q}}$ can be obtained, we can replace the structure matrix $\mathbf{A}$ in (1) with $\mathbf{Q}$ or $\hat{\mathbf{Q}}$ to have $\boldsymbol{X} = \mathbf{Q}\boldsymbol{Z}$ or $\boldsymbol{X} = \hat{\mathbf{Q}}\, \hat{\boldsymbol{Z}}$, depending on the availability $\mathbf{Q}$ or $\hat{\mathbf{Q}}$, so that, because $n = m$, the unobserved $\boldsymbol{Z}$

$$\boldsymbol{Z}=\mathbf{Q}^t\boldsymbol{X} \text{ or } \widehat{Z}=(\widehat{\mathbf{Q}})^t\boldsymbol{X}$$

can be obtained. The variance-covariance matrix of $\boldsymbol{Z}$ is now given by matrix $\mathbf{\Lambda}$, which contains eigenvalues of $\mathbf{\Sigma}$ on the diagonal, i.e.,

$$V(\boldsymbol{Z})=\varepsilon(\boldsymbol{Z}-\varepsilon(\boldsymbol{Z}))(\boldsymbol{Z}-\varepsilon(\boldsymbol{Z}))^{t}=\mathbf{Q}^{t}\times\varepsilon(\boldsymbol{X}-\varepsilon(\boldsymbol{X}))(\boldsymbol{X}-\varepsilon(\boldsymbol{X}))^{t}\times\mathbf{Q}=\boldsymbol{\Lambda}$$

$$=\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \dots & 0\\ . & . & \dots & .\\ 0 & 0 & \dots & \lambda_n\end{bmatrix}=\begin{bmatrix}V(Z_1) & Cov(Z_1,Z_2) & \dots & Cov(Z_1,Z_n)\\ Cov(Z_2,Z_1) & V(Z_2) & \dots & Cov(Z_2,Z_n)\\ . & . & \dots & .\\ Cov(Z_n,Z_1) & Cov(Z_n,Z_2) & \dots & V(Z_n)\end{bmatrix}$$

with the implication that the new transformed variables

$$Z_1=\boldsymbol{Q}^{t}{}_{\mathbf{1}}\boldsymbol{X},Z_2=\boldsymbol{Q}^{t}{}_{\mathbf{2}}\boldsymbol{X},\dots,Z_n=\boldsymbol{Q}^{t}{}_{\boldsymbol{n}}\boldsymbol{X}$$

are uncorrelated and have the respective eigenvalues of $V(\mathbf{X}) = \boldsymbol{\Sigma}$ as their variances, i.e.,

$$V(Z_i)=\lambda_i \text{ and } Cov(Z_i,Z_j)=0, \text{ for all } i,\ j=1,2,\dots,n \text{ and } i\neq j$$

Note that sums of the variances of $Z_1, Z_2,\dots, Z_n$ and $X_1, X_2,\dots, X_n$ are identical:

$$\mathrm{tr}\left(\sum\right)=\sum_{k=1}^{n}V(X_k)=\mathrm{tr}\left(\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{t}\right)=\mathrm{tr}\left(\mathbf{Q}^{t}\mathbf{Q}\boldsymbol{\Lambda}\right)=\mathrm{tr}\left(\boldsymbol{\Lambda}\right)=\sum_{k=1}^{n}V(Z_k)$$

From a geometric point of view, metric properties of both the transformed variables $Z_1, Z_2,\dots, Z_n$ and the original variables $X_1, X_2,\dots, X_n$ are identical. For instance, $|\boldsymbol{Z}| =(\mathbf{X}^{t}\mathbf{Q}^{t}\mathbf{Q}\mathbf{X})^{1/2}= |\boldsymbol{X}|$, which means the vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ have identical lengths. Furthermore, inner products in both spaces are identical: $\boldsymbol{X}_i^{t}\boldsymbol{X}_j = (\mathbf{Q}\boldsymbol{Z}_i)^{t}(\mathbf{Q}\boldsymbol{Z}_j)= \boldsymbol{Z}_i^{t}\mathbf{Q}^{t}\mathbf{Q}\boldsymbol{Z}_j = \boldsymbol{Z}_i^{t}\boldsymbol{Z}_j$. Hence, both $\boldsymbol{X}_i^{t}\boldsymbol{X}_j = \boldsymbol{Z}_i^{t}\boldsymbol{Z}_j$ and $|\boldsymbol{Z}| = |\boldsymbol{X}|$ imply that the pairs $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ and $(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ have the same positions with respect to each other in both the original and transformed spaces because $\boldsymbol{X}_i^{t}\boldsymbol{X}_j = \cos\theta\times|\boldsymbol{X}_i|\times|\boldsymbol{X}_j|$ and $\boldsymbol{Z}_i^{t}\boldsymbol{Z}_j = \cos\rho\times|\boldsymbol{Z}_i|\times|\boldsymbol{Z}_j|$ result in $\theta = \rho$. Similarly, the corresponding areas and volumes are identical in both spaces. Because multiplication from the left of the vector $\boldsymbol{X}$ by an orthogonal matrix $\mathbf{Q}^{t}$ or multiplication of $\boldsymbol{Z}$ by the orthogonal $\mathbf{Q}$ yields rotations of these vectors, the original $\boldsymbol{X}$ and the transformed $\boldsymbol{Z}$ are but rotated vectors with all other metric properties being left intact (i.e., they are so-called rigid motions of each other). For that reason, PCA is sometimes presented as a simple rotation. All of the above discussion applies to sample counterparts that result from replacing $\mathbf{A}$ with $\hat{\mathbf{Q}}$ except, of course, for relevant statistical inferential issues.

If these eigenvalues are ranked as $\lambda_{(n)} \leq \lambda_{(n-1)} \leq\dots\leq \lambda_{(2)} \leq \lambda_{(1)}$ and the corresponding eigenvectors $\boldsymbol{Q}_{(n)}, \boldsymbol{Q}_{(n-1),}\dots, \boldsymbol{Q}_{(1)}$ are ordered accordingly, then the uncorrelated transformed variables $Z_{(n)}= \boldsymbol{Q}^{t}{}_{(n)}\boldsymbol{X}$, $Z_{(n-1)} = \boldsymbol{Q}^{t}{}_{(n-1)}\boldsymbol{X},\dots, Z_{(1)} = \boldsymbol{Q}^{t}{}_{(1)}\boldsymbol{X}$ can be ordered in terms of the magnitude of their variances. The same ordering also holds for the factorized matrices

$$\sum\nolimits_{(k)}=\lambda_{(k)}\mathbf{Q}_{(k)}\mathbf{Q}_{(k)}{}^{t}$$

which can be ordered in a decreasing fashion for k=1,2,…,n

$$\sum = \sum_{(1)} + \sum_{(2)} + \ldots + \sum_{(n-1)} + \sum_{(n)}$$

Practically, this means that some reduction of dimension can be obtained on the basis of the magnitude of decomposed variances by cutting out small-ordered ones as negligible. For instance, the number of significant $Z_{(i)}$'s can be restricted to the first k variables, k < n, satisfying $\sum_{i=1}^{k} \lambda_{(i)} = \delta$ for some predetermined δ level of variance. There are some inferential issues on determination of the number *k* of significant variables (Anderson, 1984, pp. 468–479).

The decomposition in (5) is sometimes referred to as eigen decomposition to distinguish it from other decompositions such as Cholesky decomposition, etc. (Anderson, 1984, p. 586; Press et. al., 1992). When matrices are defined over complex numbers, orthogonal matrices become unitary matrices and their transposes naturally are conjugate transposes of these matrices. Because the matrix **Σ** is symmetric, resulting eigenvalues will always be real.

The formal setup of the PCA analysis is based on a structural model such as the model (1) above. Therefore, its validity depends on the validity of the model as compared with the real phenomenon that it models. For instance, data that do not involve a location parameter and/or a scale factor are hardly suitable for a framework like (1). Also, as an alternative to a complex unmanageable system based on $X_1, X_2, \ldots, X_n$, the analysis aims at obtaining some system based on variables $Z_1, Z_2, \ldots, Z_k$ that are smaller in number ($k < n$), simpler in nature, and bear the same information (eigenvalues). However, this is achieved at the expense of orthogonality and/or un-correlatedness restrictions imposed on $Z_1, Z_2, \ldots, Z_k$. These impositions may not be realistic for some applications areas such as biological systems. Application of the technique is restricted to phenomena with distributions tracing non-spherical contours in their domains because little will be gained in the spherical case. Finally, the mathematical tool on which standard PCA is based applies to nonsingular decompositions, which requires that $n = m$ in the model (1). Most real-life phenomena, however, seem to present a singular structure. The last restriction is alleviated by the next technique in the sequel: SVD. PCA does not seem to have lost its attractiveness despite prevailing high-dimensionality problems observed recently (Hastie et al., 2000). It is interesting further to note that PCA can be used for such a theoretical issue as construction of bivariate distributions (Gurrera, 2005). The use of $\hat{\mathbf{Q}}$ for **A** invites further inferential issues such as sampling distributions of eigenvalues and eigenvectors of $\hat{\boldsymbol{\Sigma}}$ (Anderson, 1984, pp.465–468, pp. 473–477, Chapter 13; Seber, 1984, pp. 35–38; Joliffe, 2002;Kendall, 1975, Chapter 2; Johnson and Wichern, 2002, Chapter 8).

Singular Value Decomposition

SVD is based on a matrix factorization technique that dates from the 19th century and was developed by several mathematicians in linear algebra and differential geometry (Stewart, 1993; Eckart. and Young, 1936). The technique warrants factorization of any real matrix in a way similar to spectral decomposition of square matrices. As such, given an ($n \times m$) rectangular matrix **C,** SVD is actually a spectral decomposition of symmetric square positive semi-definite matrices $\mathbf{C}^t\mathbf{C}$ and $\mathbf{C}\mathbf{C}^t$. For any ($n \times m$) rectangular matrix **C,** the products $\mathbf{C}^t\mathbf{C}$ and $\mathbf{C}\mathbf{C}^t$ are known to be symmetric and positive semi-definite (Scheffe, 1959, p. 399) with real non-negative eigenvalues, and these products are thus suitable for spectral decomposition explained above in connection with PCA. Assume hence that an ($n \times m$) matrix **C** is factored out the way that the matrix **Σ** is decomposed as in (5) above with **Q** now being represented by **U** when it is positioned on the left of the diagonal matrix and $\mathbf{Q}^t$ by **V** when it is on the right:

$$\mathbf{C}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{t}}=\sum_{i=1}^{\min\{n,m\}}\lambda_i \boldsymbol{U}_{\mathrm{i}}\boldsymbol{V}^{\mathrm{t}}{}_{\mathrm{i}} \tag{7}$$

where $\boldsymbol{U}_{\mathrm{i}}$'s and $\boldsymbol{V}_{\mathrm{i}}$'s are orthonormal columns of **U** and **V**, respectively, and **D** is diagonal with real nonnegative entries λ_i's. The factorization (7) is sometimes interpreted as **C** being orthogonally (unitarily) equivalent or similar to the diagonal matrix **D**. There are various modes of SVD for the dimensions of the matrices involved. For instance, (i) the matrix **U** can be ($n\times n$) with n orthonormal (unitary) columns $\boldsymbol{U}_i$, called left singular (eigen array) vectors, and **V** is an ($m\times m$) matrix with m orthonormal columns $\boldsymbol{V}_i$, called right singular vectors (eigen genes in the biological example considered above) such that

$$\mathbf{U}^{\mathrm{t}}\mathbf{U}=\mathbf{I}_{\mathrm{n}} \text{ and } \mathbf{V}^{\mathbf{t}}\mathbf{V}=\mathbf{I}_{\mathrm{m}}$$

D is an ($n\times r$), $r = \min\{n,m\}$, diagonal matrix with positive diagonal entries called singular values (eigen expressions) of **C**. (ii) A second mode corresponds to the case where the dimensions of the matrix **U** are ($n\times m$) with n orthonormal left singular vectors and the ($m\times m$) matrix **V** itself is orthogonal with m right singular vectors, such that

$$\mathbf{U}^{\mathrm{t}}\mathbf{U}=\mathbf{V}^{\mathbf{t}}\mathbf{V}=\mathbf{I}_{\mathrm{m}}$$

Hence, **D** becomes a ($m\times m$) diagonal matrix with non-negative real singular values of **C** on its diagonal. The thin, compact, and truncated types of SVD are not discussed here because of space considerations.

To explain technicalities in SVD, we denote the real non-negative eigenvalues of the matrices $\mathbf{C}^{\mathrm{t}}\mathbf{C}$ and $\mathbf{C}\mathbf{C}^{\mathrm{t}}$ with $\kappa_1, \kappa_2, \ldots, \kappa_{\mathrm{r}}$ with r being $\min\{n,m\}$. Given the decomposition $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{t}}$ as in (7) with dimensions given for instance as in (i) above, we obtain

$$\mathbf{C}^{\mathrm{t}}\mathbf{C}=\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{t}}\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{t}}=\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{t}}$$

from which, by multiplication with **V** on the right, we have

$$\mathbf{C}^{\mathrm{t}}\mathbf{C}\,\mathbf{V}=\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{t}}\mathbf{V}=\mathbf{V}\mathbf{D}^2 \tag{8}$$

Setting $\mathbf{G}= \mathbf{C}^{\mathrm{t}}\mathbf{C}$ and noting that $\mathbf{V}= [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_m]$ and $\mathbf{D}^2 = \mathrm{diag}\,(\kappa_1, \kappa_2, \ldots, \kappa_{\mathrm{r}})$, (8) yields

$$\mathbf{G}\,\boldsymbol{V}_i=\kappa_{\mathrm{i}}\,\boldsymbol{V}_{\mathrm{i}},\ \ i=1,2,\ldots,r$$

which suggests that eigenvectors $\boldsymbol{V}_{\mathrm{i}}$ of the matrix $\mathbf{G} = \mathbf{C}^{\mathrm{t}}\mathbf{C}$ are the right singular vectors of the matrix **G,** and the singular values of **C** are given by absolute square roots of the corresponding eigenvalues of $\mathbf{G} = \mathbf{C}^{\mathrm{t}}\mathbf{C}$**.** When these non-negative singular values are ranked in terms of their descending magnitudes, i.e., $\sqrt{\kappa_{(r)}} \leq \sqrt{\kappa_{(r-1)}}\,!\ddot{\mathrm{U}}\cdots!\ddot{\mathrm{U}}\,\sqrt{\kappa_{(1)}}$, then corresponding singular

vectors can also be ordered in descending degree of importance such as ,…,$\boldsymbol{U}_{(1)}$,…, $\boldsymbol{U}_{(r-1)}$, $\boldsymbol{U}_{(r)}$ and $\boldsymbol{V}_{(1)}$,…,$\boldsymbol{V}_{(r-1)}$, $\boldsymbol{V}_{(r)}$. For clarity, matrix **D** can be rearranged, i.e., for n < m and m<n, giving respectively

$$\mathbf{D}=\begin{bmatrix} \sqrt{\kappa_{(1)}} & 0 & \dots & 0 & \vdots & 0 & \dots & 0 \\ 0 & \sqrt{\kappa_{(2)}} & \dots & 0 & \vdots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\kappa_{(n)}} & \vdots & 0 & \dots & 0 \end{bmatrix} \text{ and } \mathbf{D}=\begin{bmatrix} \sqrt{\kappa_{(1)}} & 0 & \dots & 0 \\ 0 & \sqrt{\kappa_{(2)}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\kappa_{(n)}} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

Note also that (7) can also be re-expressed as

$$\mathbf{C}=\mathbf{U}\mathbf{D}\mathbf{V}^{t}=\sum_{i=1}^{\min\{n,m\}}\sqrt{\kappa_{(i)}}\,\boldsymbol{U}_{(i)}\boldsymbol{V}_{(i)}^{t}=\sum_{i=1}^{\min\{n,m\}}\mathbf{C}_{(i)},\quad \mathbf{C}_{(i)}=\sqrt{\kappa_{(i)}}\,\boldsymbol{U}_{(i)}\,\mathbf{V}_{(i)}^{t} \tag{9}$$

with $\mathbf{C}_{(i)}\,\mathbf{C}^{t}_{(j)}=\mathbf{0}$ for i ≠ j, i.e., the matrices $\mathbf{C}_{(i)}$ and $\mathbf{C}_{(j)}$ are orthogonal and hence matrix **C** can be decomposed in an additive way as well. The matrices $\mathbf{C}_{(i)}$ in (9) are uncorrelated "modes" of the original matrix **C**. Because of the ordering of singular values, the modes ($\mathbf{C}_{(i)}$'s) in (9) are also ordered, so that $\mathbf{C}_{(i)}$'s corresponding to negligible (small) singular values are sometimes loosely defined as noise and can thus be ignored in the analysis.

Simple arithmetic reveals, on the other hand, that the eigenvalues and eigenvectors of the matrix $\mathbf{F}=\mathbf{C}\mathbf{C}^{t}$ will now be $\kappa_1, \kappa_2,\dots, \kappa_m$ and $\boldsymbol{U}_1,\boldsymbol{U}_2,\dots,\boldsymbol{U}_m$, respectively. Accordingly, by definition of these values and vectors, we will have

$$\mathbf{F}\,\boldsymbol{U}_i=\kappa_i\,\boldsymbol{U}_i,\quad i=1,2,\dots,m$$

The decomposition will then be

$$\mathbf{C}=\mathbf{U}\mathbf{D}\mathbf{V}^{t}=\sum_{i=1}^{m}\sqrt{\kappa_{(i)}}\,\boldsymbol{U}_{(i)}\boldsymbol{V}_{(i)}^{t}=\sum_{i=1}^{m}\boldsymbol{Y}_{(i)},\quad \boldsymbol{Y}_{(i)}=\sqrt{\kappa_{(i)}}\,\boldsymbol{U}_{(i)}\,\boldsymbol{V}_{(i)}^{t} \tag{9}$$

Note that the matrices $\mathbf{Y}_{(i)}$ of the decomposition (9) are orthogonal in the sense that for all $i \neq j$, we have

$$\boldsymbol{Y}_{(i)}\boldsymbol{Y}^{t}_{(j)}=\sqrt{\kappa_i\kappa_j}\,\boldsymbol{U}_{(i)}\boldsymbol{V}_{(i)}^{t}\,\boldsymbol{V}_{(j)}\,\boldsymbol{U}_{(j)}^{t}=\mathbf{0}$$

Furthermore, as it is the case with spectral decomposition in PCA, it is possible to obtain a pseudo (Monroe-Penrose) inverse $\mathbf{C}^{-}$ of **C** as

$$\mathbf{C}^{-}=\mathbf{U}\mathbf{D}^{-}\mathbf{V}^{t}$$

where $\mathbf{D}^{-}$ is the corresponding pseudoinverse of **D.** Because **D** is a diagonal matrix, its pseudoinverse will be also diagonal and will correspond thus to its transposition. So that, for the matrix **D** given earlier, i.e.,

$$\mathbf{D}=\begin{bmatrix} \sqrt{\kappa_{(1)}} & 0 & \dots & 0 \\ 0 & \sqrt{\kappa_{(2)}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\kappa_{(n)}} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

the generalized inverse looks like

$$\mathbf{D}^{-}=\begin{bmatrix} a_{(1)} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & a_{(2)} & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{(n)} & 0 & 0 & \dots & 0 \end{bmatrix}=[\mathbf{D^{-1}}, \mathbf{0}]$$

where $a_{(i)}=\frac{1}{\sqrt{\kappa_{(i)}}}$, i = 1,2,…,n.

This much SVD algebra is suffient for this brief review. There are various applications of SVD in different areas of statistical analysis, but its main use is in general regression analyses yielded by

$$\min_{\mathbf{A},\mathbf{Z}} \|\mathbf{X}-\mathbf{A}\mathbf{Z}\|^2$$

where the noiseless version of the model (3) above is considered and the norm is the Frobenius (Euclidean) matrix norm defined for any rectangular matrix **M** as

$$\|\mathbf{M}\|=\sqrt{\ddagger\text{”}\ddagger\text{”}|m_{ij}|^2}|=\sqrt{trace(\mathbf{M}^t\mathbf{M})}$$

This particular use shows how SVD was first used (Lawson and Hanson, 1995).

Another use of SVD in the sense of three-matrix factorization as in (7) relates to decomposition of the observation matrix **X** and is usually observed in biology, where the matrix **U** is used for bioassay, elements of the matrix **D** are called singular values, and elements of the matrix **V** are interpreted as eigen genes. Matrix **X** is broken thus down to orthogonal summands, i.e.,

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{t}}=\sum_{i=1} d_{(i)}\ \boldsymbol{U}_{(i)}\boldsymbol{V}_{(i)}{}^{\mathrm{t}}=\sum_{i=1}\mathbf{X}_{(i)},\ \ \mathbf{X}_{(i)}=d_i\ \mathbf{U}_{(i)}\mathbf{V}_{(i)}{}^{t}$$

Matrix **U** is thus functionally identical to matrix **A** of the structural models in (1) and (2).

One of the main uses of SVD in the current context is for decomposition of covariance matrices. Because the covariance matrix of two differing vectors of distinct dimensions is rectangular, it is decomposable in the sence of SVD. Let a pair of vectors such as the (n×1) vector $\boldsymbol{X}$ and the (m×1) vector $\boldsymbol{Y}$ be given. Their covariance

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=\mathbf{M}=\varepsilon(\boldsymbol{X}-\varepsilon(\boldsymbol{X}))(\boldsymbol{Y}-\varepsilon(\boldsymbol{Y}))^{\mathrm{t}}$$

is an (n×m) rectangular matrix and is therefore decomposable singularly as

$$\mathbf{M}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{t}}$$

Consequently, as in PCA, the variance-covariance matrix of any transformed vector $\boldsymbol{Z}=\mathbf{U}^{\mathrm{t}}\boldsymbol{X}$ will then be

$$\varepsilon(\boldsymbol{Z}-\varepsilon(\boldsymbol{Z}))(\boldsymbol{Z}-\varepsilon(\boldsymbol{Z}))^{\mathrm{t}}=\mathbf{U}^{\mathrm{t}}\mathrm{V}(\boldsymbol{X})\mathbf{U}=\mathbf{U}^{\mathrm{t}}(\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{t}})\mathbf{U}=\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{\mathrm{t}}=\boldsymbol{\Lambda}^{*}$$

where

$$\boldsymbol{\Lambda}^{*}=\begin{bmatrix}\lambda_{(1)} & 0 & \dots & 0\\ 0 & \lambda_{(2)} & \dots & 0\\ \vdots & \vdots & \dots & \vdots\\ 0 & 0 & \dots & \lambda_{(\mathrm{r})}\end{bmatrix}\ \text{and}\ \boldsymbol{\Phi}=\mathbf{U}^{\mathrm{t}}\mathbf{Q}$$

with r = min {n,m}. We then use the spectral decomposition $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{t}}$ for V($\boldsymbol{X}$), and in the last equality, we utilize the fact that multiplication of two orthogonal matrices such as $\mathbf{U}^{\mathrm{t}}$ and **Q** yields again an (m×n) matrix **Φ** with orthonormal columns:

$$\boldsymbol{\Phi}=\mathbf{U}^{\mathrm{t}}\mathbf{Q}=\begin{pmatrix}U^{t}_{(1)}\\ U^{t}_{(2)}\\ \vdots\\ U^{t}_{(m)}\end{pmatrix}(Q_{(1)},Q_{(2)},\cdots,Q_{(n)})=\begin{bmatrix}U^{t}_{(1)}Q_{(1)} & U^{t}_{(1)}Q_{(2)} & \dots & U^{t}_{(1)}Q_{(n)}\\ U^{t}_{(2)}Q_{(1)} & U^{t}_{(2)}Q_{(2)} & \dots & U^{t}_{(2)}Q_{(n)}\\ \vdots & \vdots & \vdots & \vdots\\ U^{t}_{(m)}Q_{(1)} & U^{t}_{(m)}Q_{(2)} & \dots & U^{t}_{(m)}Q_{(n)}\end{bmatrix}$$

so that a typical element of $\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{\mathrm{t}}$ becomes

$$\boldsymbol{U}^{\mathrm{t}}{}_{(\mathrm{i})}[\sum_{k=1}^{n}\lambda_{(k)}\,\boldsymbol{Q}_{(\mathrm{k})}\boldsymbol{Q}^{\mathrm{t}}{}_{(\mathrm{k})}]\boldsymbol{U}_{(\mathrm{j})}=\begin{cases}\lambda_{(k)}, & i=j=k\\ 0, & i\neq j.\end{cases}$$

The latter conclusion is obviously identical to the result that can be obtained from PCA.

In sum, SVD is more general than PCA because it is applicable to cases where PCA is applied but the converse is not valid. Furthermore, as mentioned earlier, an appropriate type of SVD (thin, compact, and truncated) can be chosen for a problem. Like PCA, SVD is also used for data reduction. As is the case with PCA, the application of SVD is restricted to phenomena that have statistical distributions with non-spherical contours. All in all, SVD has the same drawbacks as PCA because it is based on the same structural approach, which may be restrictive in some application areas. Also, the orthogonality restriction imposed by the analysis may not be suitable for some types of data. An interesting illustrative application is in Chung and Seabrook (2004). The next approach that we discuss goes to a further extreme in the last direction.

Independent Component Analysis (Blind Source Separation)

ICA seems to have been proposed by several authors from various disciplines in the last quarter of a century (Comon, 1994). However, the main areas of origin are signal processing and neural network analyses. ICA is also known as blind (myopic) source separation. The qualifiers “blind” or “myopic” are used to point out the property that the sole observable of the system is its output, i.e., the vector $\boldsymbol{X}$ of the model (1). “Independence” is affixed to the name, because of technical reasons concerning input, i.e., the vector $\boldsymbol{Z}$ of the model can satisfy certain contrast functions in signal processing when its coordinates are independent. The name “independent component analysis” seems to have been suggested by Jutten and Herault (1991) with respect to “principal component analysis”; they also refer to it as “blind source separation”.

In model (1) and/or (2) the variables $X_1, X_{2,}\ldots, X_n$ represent random observable variables and the variables $Z_1, Z_{2,}\ldots, Z_m$ represent unobservable (latent) variables with **A** being a mixing matrix. When the matrix **A** is orthogonal as in PCA and SVD, the variables $Z_1, Z_{2,}\ldots, Z_m$ become orthogonal transformations (projections) of the variables $X_1, X_{2,}\ldots, X_n$. The orthogonality property results in mutual uncorrelatedness of $Z_1, Z_{2,}\ldots, Z_m$ (see covariance matrix $\boldsymbol{\Lambda}$ above). Uncorrelatedness does not imply independence, which is obviously a stronger requirement than orthogonality. In terms of the method-of-moments terminology, orthogonality relates only to second moments (or cumulants) whereas independence also involves higher moments (or cumulants). ICA is based on the assumption that the generating sources for the variables $X_1, X_{2,}\ldots, X_n$ are independent. The joint distribution of $X_1, X_{2,}\ldots, X_n$ may be known or unknown. The objective of the analysis is to determine an unknown **A** and the corresponding unobservable vector $\boldsymbol{Z}=(Z_1, Z_{2,}\ldots, Z_m)^t$ through some observations on $\boldsymbol{X} = (X_1, X_{2,}\ldots, X_n)^t$.

Gaussian distributions are not appropriate for that objective, because, **A** cannot be identified and is therefore not estimable by these distributions. In fact, when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two vectors with each being an orthogonal transformation of the other, their metrics $|\boldsymbol{X}|$ and $|\boldsymbol{Y}|$ will thus be identical, i.e., $|\boldsymbol{X}| = |\boldsymbol{Y}|$. When both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed further to be Gaussian distributions with identical means and variances (for instance, when they have zero mean vectors and when their variances are equal to identity matrices of the same dimension), their densities will look like

$$k\,exp\{-\frac{1}{2}|\boldsymbol{X}|^2\}=k\,exp\{-\frac{1}{2}|\boldsymbol{Y}|^2\}$$

Setting $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z}$ and $\boldsymbol{Y} = \mathbf{A}\boldsymbol{Z}$ as in the model (1),

$$k\,exp\{-\frac{1}{2}|\mathbf{A}\boldsymbol{Z}|^2\}=k\,exp\{-\frac{1}{2}|\mathbf{A}\boldsymbol{Z}|^2\}$$

which shows that neither **A** nor ***Z*** can be identified by these densities.

Furthermore, Gaussian distributions are proved to have the largest entropy among the distributions of random variables with identical means and variances (Papoulis and Pillai, 2002, p.669). If we set $H(\boldsymbol{X})$ for the entropy of a random vector ***X*** with coordinates X_i's and *H(S)* stands for a system *S* with components X_i's, the entropy corresponding to the whole vector ***X*** or to the system *S* is then given as

$$H(\boldsymbol{X})=\left\{\begin{array}{c}-\sum_{i=1}^{\infty}P(X=x_i)\times\log P(X=x_i)\\ -f(x)\log f(y)dx\end{array}\right\}=H(S)$$

where *P* stands for the joint probability density function (pdf) for the discrete ***X, f*** denotes the joint pdf in the continuous case, and the logarithm is with respect to *2* or *e* base. The concept of entropy discussed here corresponds to Schroedinger's (1944, p.73) concept of negative entropy. As such, it is an information measure. Thus, if $H(\boldsymbol{X_{Gaus}})$ stands for the entropy of a random vector $\boldsymbol{X_{Gaus}}$, which has a Gaussian distribution, and *H(**X**)* denotes the entropy of a distribution of non-Gaussian random variables represented by the vector ***X***, then the non-negative magnitude

$$J(\boldsymbol{X})=H(\boldsymbol{X}_{Gaus})-H(\boldsymbol{X})$$

yields the discrepancy of entropies between the Gaussian-distributed $\boldsymbol{X_{Gaus}}$ and any other ***X*** supposed to be different from $\boldsymbol{X_{Gaus}}$. Accordingly, the larger $J(\boldsymbol{X})$ is the more different $\boldsymbol{X_{Gaus}}$ will be from ***X***. One method for checking non-normality of the distribution involved consists of maximizing $J(\mathbf{X}) = J((X_1, X_2, \ldots, X_n)$. Obviously, $J(\boldsymbol{X})$ is a variation of negentropy (not to be confused with Schroedinger's concept), which is shown to be invariant under linear transformations. The negentropy function $J(\boldsymbol{X})$ can be considered as the initial technical process for estimating matrix **A** in the model (1) on the basis of joint distribution of $X_1, X_2, \ldots, X_n$. However, if this estimation requires prior knowledge on the relevant pdf or its estimate, then this technique will be computationally difficult.

Before matrix **A** is estimated, statistically independent components of or statistically independent sources for the rows (or columns) of the matrix **X** in (1) must be checked (Lee et. al., 1998). A method used for that purpose is based on minimization of the mutual information function of Papoulis and Pillai (2002, p.647) or the transmission function of Conant (1968, p. 11) defined as

$$T(\boldsymbol{X})=\sum_{i=1}^{n}H(X_i)-H(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}H(X_i)-H(\boldsymbol{X}) \quad (10)$$

where $H(X_i)$ is the entropy of each individual random variable X_i of an *n*-dimensional column vector $\boldsymbol{X} = (X_1, X_2,\ldots,X_n)^t$. $T(\boldsymbol{X})$ is a non-negative quantity indicating discrepancy between entropies of the dependence case, i.e., $H(X_1, X_2,\ldots,X_n)$, and the *n* independence case, i.e.,

$\sum_{i=1}^{n} H(X_i)$; the latter is always larger than the former (Papoulis and Pillai, 2002, p.646). (The proof given by these authors corresponds to the two-subsystem [two-variable] case but can be easily extended to the *n*-variable case as well.) The larger the quantity $T(\boldsymbol{X})$ in (10), the more dependent $X_1, X_2,\ldots,X_n$ become, in which case $X_1, X_2,\ldots,X_n$ are not appropriate for estimating matrix **A**. Obviously, to be able to minimize $T(\boldsymbol{X})$, marginal and joint distributions of $X_1, X_2, \ldots,X_n$ must be known and, as previously remarked, estimation of pdf values is difficult. Methodological developments by Conant (1968 and 1972) allow all computations involved in $T(\boldsymbol{X})$ to be easily based on empirical frequencies.

The essential method for estimating matrix **A** in (1) is a maximum likelihood technique that is applicable when the relevant densities are known (Hyvarinen and Oja, 2000;Comon, 1994;Hyvarinen et al., 2001). Before the application of this method, to be on the safer side, the required non-normality of the distribution involved and the necessary independence conditions used for estimation must be checked through the two methods previously discussed.

In sum, ICA has a well-founded statistical basis. The analysis displays the same drawbacks as PCA and SVD because its foundation is a structural approach that may not fit the data of some areas of the application. ICA evidently eliminates the orthogonality requirement of PCA and SVD by introducing the independence condition, but the latter actually is stricter than the former (for an illustrative comparative review of the above three techniques, see, Srinivasan et. Al.). Application of this technique seems to be restricted to non-Gaussian distributions. All three previous decomposition techniques are designed to choose matrix **A** of model (1) from among the matrices that satisfy orthogonality and independence properties imposed on real data. Such matrices may not be appropriate for biological data. We have thus another technique as proposed by the decomposition approach: NCA.

Network Component Analysis

Some recent discussions (Liao et. al., 2003) concerning PCA, SVD, and ICA note that the hypothesized relationship between $\boldsymbol{Z}$ and $\boldsymbol{X}$ (assumed to exist by these three techniques for the models (1) or (3)) imposes some statistical constraints on data; for example, orthogonality (orthonormality) and independence of unobservable variables are not warranted in some applications, such as biological experimentation. These methods obtain a (random) value of matrix **A** from a class of all real (complex)-valued matrices satisfying the hypothesized relationship. In an actual case, $\boldsymbol{X}$ is maintained to be produced by $\boldsymbol{Z}$ according to a bipartite network system between two distinct groups of phenomena such as phenotype and genotype (Figure 1).

The ($n \times N$) observation matrix **X** on the left side of (3) can be reconstructed as

$$\mathbf{X}=\mathbf{BC} \tag{11}$$

where the ($m \times N$) matrix **C,** which replaces matrix **Z** in (3) above and is now called '*regularity matrix*', consists of *N* samples of *m* signal inputs (genotypes) with the condition $m < N$. The ($n \times m$) *connectivity matrix* **B,** which replaces matrix **A** in (3) above, indicates the number of ways by which input signals (genotypes) are connected to signal outputs (phenotypes) in **X** so that each input may or may not be connected to each output. Thus, a typical entry (b_{ij}) of **B** shows how many times the j^{th} signal input is connected to the i^{th} output; it is zero when there is no connectivity between the relevant inputs and outputs. Regularity matrix **C** is assumed to be of full row rank (i.e., $r(\mathbf{C}) = m$); connectivity matrix **B** is of full column rank (i.e., $r(\mathbf{B}) = m$) with each column of **B** containing at at least ($m - 1$) zeros. By (3) and (10), we have

$$\mathbf{X}=\mathbf{AZ}=\mathbf{BC} \tag{12}$$

As follows from (12), the statistical assumptions concerning orthogonality or independence on the constituent factors ***Z*** work through the regularity matrix **C** on the right side of (11) in such a way as to make **C** comply with these orthogonality and/or independence criteria. The decompositions corresponding to the right side of (11) are unique up to a nonsingular transformation **P** such as $\mathbf{A}^*=\mathbf{AP}$; $\mathbf{Z}^*=\mathbf{P}^{-1}\mathbf{Z}$, $\mathbf{B}^*=\mathbf{BP}$ and $\mathbf{C}^*=\mathbf{P}^{-1}\mathbf{C}$, which yields

$$\mathbf{X}=\mathbf{A}^*\mathbf{Z}^*=\mathbf{AP}\,\mathbf{P}^{-1}\mathbf{Z}=\mathbf{AZ} \text{ and } \mathbf{B}^*\mathbf{C}^*=\mathbf{BPP}^{-1}\mathbf{C}=\mathbf{BC}=\mathbf{X}$$

As is proven in Liao et al. (2003, Appendix 1), when on the conditions **C** is of full row rank, the rank of **B** equals the number of its columns, and each column of **B** contains at least (m − 1) zeros, the transformation matrix **P** can only be a diagonal matrix so that the decomposition **X** = **BC** is unique up to a diagonal scalar. This ensures identifiability of a system up to diagonal scalar in the sense that, corresponding to certain regularity **C,** only one connectivity (network) matrix **B** produces the observations **X.**

From a computational standpoint, the matrix decomposition **BC** of **X** is obtained through an iterative optimization (minimization with respect to matrices **B a**nd **C**), an algorithm applied to the square of Frobenius (Euclidean) matrix norm mentioned earlier, e.g.,

$$\|\mathbf{M}\|=\sqrt{\ddagger''\ddagger''|m_{ij}|^2|}$$

of the rectangular matrix **M** = **X** − **BC**. Hence, **B** and **C** are estimated from **X** by using a two-step least-squares algorithm using the objective function

$$\min_{\mathbf{BC}}\|\mathbf{X}-\mathbf{BC}\|^2$$

subject to the side condition that the connectivity pattern **B** belongs to a manifold of matrices with a certain pattern (e.g., entries representing connectivity are arbitrary non-zero natural numbers initially and non-connectivity is represented by zero). The detailed algorithm of estimation itself is summarized and can be found in Liao et al. (2003, Appendix 2).

To sum up, NCA has fundamentally the same foundation as the previous analyses: It has a structural and therefore a parametric approach, the only difference being the estimate of matrix **A** from a narrower class of matrices satisfying a certain network configuration. It therefore displays the same drawbacks as PCA, SVD, and ICA.

Decomposition by Information Transfer

As it is the case with the previous approaches, the idea behind the decomposition by ITF is to find a much simpler system than a given system, simplicity being defined as lack of complexity of systems (especially for hierarchical systems that are composed of some layers of subsystems). Unlike PCA, SVD, ICA, and NCA, ITF does not specify a structural model and analysis is not carried out in terms of such a model. Ignoring the specific natures of components

that make up a system, a measure of complexity of a system can be related to the earlier-defined entropy $H(S)$ and transmission $T(S)$ functions for a system $S = \{X_1, X_2,\ldots,X_n\}$ with n ordered components $X_1, X_2,\ldots,X_n$.. For technical reasons, S is not necessarily a vector with coordinates $X_1, X_2,\ldots,X_n$ but represents a system with ordered components in the sense that, unless stated otherwise, all permutations of $\{X_1, X_2,\ldots,X_n\}$ stand for distinct systems. Thus, because complexity is related to information (actually, a lack of information), the following non-negative magnitude, which is in fact a measure of information,

$$\sum_{i=1}^{n} H(X_i)$$

can be used to measure lack of complexity of a system when n individuals (components) are taken one by one and their inherent potentials (measured in terms of entropy) are added up arithmetically. The underlying assumption for such measure is that individuals act independently and do not exhibit a coherent body. Similarly,

$$H(X_1, X_2, \ldots, X_n)=H(S)$$

is also a measure of information (lack of complexity) when interactive behavior of n components is taken into account, i.e., when individuals form a coherent body of a system.

The transfer function $T(\boldsymbol{X})$ already defined as a difference of the two foregoing measures of information in (10) is now designated by $T(S, \boldsymbol{p})$ and will be defined in terms of the components of S:

$$T(S, \boldsymbol{P})=\sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n) \tag{13}$$

This obviously is a non-negative measure of change in information between a system composed of a random collection of n independent individuals and a system composed of an integrated body of these n individuals. In a way, the transfer function indicates useable information that exists in components for the system. The larger the magnitude, the more remote the system S will be from the chaotic case of n components being unable to form a system. The symbol $\boldsymbol{p}$, expressed precisely as $\boldsymbol{p}$ $\{X_1, X_2,\ldots,X_n\}$, represents a partition of S. The notation $T(S, \boldsymbol{p})$ hence emphasizes that information transformation depends on the given system S as well as the specific partition $\boldsymbol{p}$ of S under consideration. Roughly, the contribution of $\boldsymbol{p}$ to $T(S, \boldsymbol{p})$ n corresponds to the sum $\sum_{i=1}^{n} H(X_i)$ on the right side of (13), and that of the S-part is given by $H(X_1, X_2,\ldots,X_n)$ on the same side of the equation. Because we now are dealing with a system rather than the vector $\boldsymbol{X}$, we now bring in the system symbol S into (10) to have the notation in (13). By a partition $\boldsymbol{p}$ $\{S_1, S_2,\ldots, S_q\}$ of S, we mean a collection of disjoint subsets $S_1, S_2, \ldots, S_q$, $(q \leq n)$ of S that exhaust S, i.e., $S=\bigcup_{i=1}^{q} S_i$. The simplest partition is $\boldsymbol{p}$ $\{X_1, X_2,\ldots, X_n\}$ as in (13), where $q = n$ with $S_i =\{X_i\}$ and is called the element partition. A trivial partition is $\boldsymbol{p}$ $\{S\}$, i.e., $\boldsymbol{p}$ $\{S\}=S$ with $q=1$ whereas numerous partitions correspond to $1< q < n$, the well known of which is the dichotomous partition $\boldsymbol{p}$ $\{S_1,S_2\}$. With partition notation $\boldsymbol{p}$ $\{S_1, S_2,\ldots, S_q\}$, the ITF becomes

$$T(S,\boldsymbol{P})=T(S_1,S_2,\ldots,S_q,\boldsymbol{P})=\sum_{i=1}^{q}H(S_i)-H(S_1,S_2,\ldots,S_q) \tag{14}$$

where $H(S_1,S_2,\ldots,S_q)$ is equal to $H(X_1, X_2,\ldots,X_n)$ because $\bigcup_{i=1}^{q} S_i=S=\{X_1, X_2, \ldots, X_n\}$ and $H(S_1), H(S_2), \ldots, H(S_q)$ are the corresponding entropies of the individual elements of the partition. Equation (14) can be interpreted as unused information in $\boldsymbol{p}$ $\{S_1, S_2,\ldots, S_q\}$ for the system S. As noted earlier, (10) and its variations as in (13) and (14) are specifically discussed in Papoulis and Pillai (2002) and Conant (1968, 1972). For applications in classical statistical inference, a parallel information theoretic approach is adopted by Kullback (1959, Chapters 1 through 3).

To emphasize the interaction between these two sets when a dichotomous partition $\boldsymbol{p}$ $\{S_1, S_2\}$ and their transmission is involved, the function T is sometimes indexed by the subscript B, i.e.,

$$T_B(S_1,S_2,\boldsymbol{P})=H(S_1)+H(S_2)-H(S_1,S_2)=T(S_1,S_2,\boldsymbol{P})$$

Similarly, to emphasize interaction within the system $S=\{X_1, X_2,\ldots,X_n\}$ with an element partition $\boldsymbol{p}$, the magnitude $T(S, \boldsymbol{p})$ can also be tagged as $T_W(S, \boldsymbol{p})$, where the subscript W stands for the interaction within the system S. Because a coherent system can only be composed of a series system, a parallel system, or a mixture of both (Barlow and Prochan, 1975, Chapter 2), then the probability corresponding to the partition can range from the probability of the former, i.e., $P(S_1, S_2) = P(S_1) \times P(S_2)$ with its entropy being $H(S_1,S_2) = H(S_1) + H(S_2)$, to the probability of the latter, i.e., $P(S_1, S_2) = min\{P(S_1), P(S_2)\}$ with the corresponding entropy of the latter being $H(S_1, S_2) = min\{H(S_1), H(S_2)\}$. We have thus a maximum value for $T(S_1,S_2, \boldsymbol{p})$ that is obtained when $H(S_1, S_2) = min\{H(S_1), H(S_2)\}$ and the minimum value of $T_B(S_1,S_2, \boldsymbol{p})$ is reached when $H(S_1,S_2) = H(S_1) + H(S_2)$, in which case $T_B(S_1,S_2, \boldsymbol{p}) = 0$. The minimum value of the transfer function indicates that S_1 and S_2 are independent in statistical terms. If we let $T^U{}_B(S_1{:}S_2)$ stand for the maximal value of $T_B(S_1,S_2, \boldsymbol{p})$,

$$0 \le T_{12}=\frac{T_B(S_1,S_2,\boldsymbol{P})}{T^U{}_B(S_1,S_2)} \le 1 \tag{15}$$

For all systems partitioned as S_1 and S_2 and having joint distributions with fixed marginal distributions for S_1 and S_2 (i.e., systems with the given marginal distributions of S_1 and S_2), T_{12} will be closer to one as the underlying sub-systems S_1 and S_2 get more and more integrated to form a whole system and will approach to zero as S_1 and S_2 become disintegrated to form separate systems. T_{12} is actually some version of the absolute value of the usual correlation coefficient when S_1 and S_2 are singletons like $S = \{X_1\}$and $S_2 = \{X_2\}$. Hence, $T_{12} = 0$ means non-relatedness of subsets S_1 and S_2, and $T_{12} = 1$ implies that S_1 and S_2 depend completely on each other. This latter aspect of transfer functions is amply emphasized in Conant (1968, 1972).

For comparison of the complexities of pairs of subsystems, Conant (1968) introduced an additional instrumental index called the *interaction measure*:

$$Q_{s_i}(S_j)=T_{s_i}(S_j)-T(S_j,\boldsymbol{P})\ i\neq j=1,2,\ldots,q \tag{16}$$

which is useful in detecting a change in the complexity of S_j when it is integrated into the subsystem S_i, so that the interdependence among the components of S_j are evaluated against their joint conditional interdependence, given the integration of the elements of S_i. $T_{S_i}(S_j, \boldsymbol{p})$ in (16) denotes the conditional transmission over $S_j = \{X_{j_1}, X_{j_2}, \ldots, X_{j_r}\}$ and given the interrelatedness of the elements of S_i, and is defined as

$$T_{s_i}(S_j) = \sum_{i=j_1}^{j_r} H_{s_i}(X_i) - H_{s_i}(X_{j_1}, X_{j_2}, \ldots, X_{j_r})$$

where

$$H_{s_i}(S_j) = H(S_i, S_j) - H(S_j)$$

Unlike the $T(.)$ transfer functions that are always positive, the interaction measure $Q(.)$ in (16) can be negative (i.e., the sets S_i and S_j interact negatively), positive (i.e., S_i and S_j interact positively), or zero (i.e., S_i and S_j are stochastically independent).

The usefulness of the information transfer function in (13) or (14) for applications of systems is two-fold. When we have no usable information or conversely when we have full information on both S and $\boldsymbol{p}$, the function $T(S, \boldsymbol{p})$ ceases to be useful. Its use becomes accentuated when we have partial information on S and/or $\boldsymbol{p}$. In fact, we may be in a position to obtain a system with some given subsystems or to derive some subsystems from a given system. The former problem is composition (integration) and latter is decomposition. In composition, $\boldsymbol{p}$ is known but we have no information on S. In decomposition, we have information on S but none about $\boldsymbol{p}$. Information on $\boldsymbol{p}$ and/or S means knowledge about the distributions involved as well; "information" as used here is either in the theoretic sense, e.g., the information transfer function, or is in the daily usage sense, i.e., knowledge. Thus, for solution of the integration problem, the information transfer function is maximized over all possible systems $\boldsymbol{\mathcal{S}} = \{S | S \bullet \boldsymbol{\mathcal{S}}\}$ that yield the given common partition $\boldsymbol{p}$. In other words, $H(S) = H(S_1, S_2, \ldots, S_q)$ or $H(X_1, X_2, \ldots, X_n)$ is minimized. Conversely, a solution for decomposition (partition $\boldsymbol{p}$) is obtained by minimizing the transfer function over all possible partitions $\boldsymbol{\pi} = \{ \boldsymbol{p}: \boldsymbol{p} \bullet \boldsymbol{\pi} \}$ of the given system S. In other words, $\sum_{i=1}^{q} H(S_i)$ or $\sum_{i=1}^{n} H(X_i)$ is minimized. Composition is not simply a reversal of decomposition because, in addition to the required knowledge on relevant partition, the so-called blueprint of composition must also be known, i.e., compatibility of all marginal distributions with the given joint distribution of S must be checked (Dall'Aglio, 1972;Butterfly et. al., 2005;Arnold et al., 1999). However, the current work is only interested in decomposition.

All foregoing discussions and indices are undoubtedly relevant when distributions are known. When such information is unavailable, empirical distributions can be used as consistent estimates. Collecting observations for this purpose is costly and difficult. This is basically true for empirical assessment of joint distributions when we are interested in obtaining estimates of multivariate distributions for large sets of components. The following illustrate the initial stages where observations on element partition and pair-wise partition of the system are available:

Illustration One

A simple empirical case corresponds to the availability of observations on each individual component. Let $S = \{Y_1, Y_2, \ldots, Y_n\}$ hence be a system with n ordered components and with the given matrix $\mathbf{X}$ of discrete observations

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ y_{\kappa_1} & y_{\kappa_2} & \cdots & y_{\kappa_n} \end{bmatrix} = (y_{ij})$$

where y_{ij} denotes the i^{th} functional value ($i=1,2,\ldots, \kappa$) taken up by the j^{th} component of the system $S = \{Y_1, Y_2, \ldots, Y_n\}$. The number (say, κ) of functional values is identical for each random component for the convenience of notation. The values y_{ij} are not necessarily identical for all components. Assume further that these values are observed with the frequencies in Table 1:

By definition, the magnitudes $\boldsymbol{m_{\circ j}}$ and m in Table 1 are $\sum_{i=1}^{\kappa} m_{ij} = \boldsymbol{m_{\circ j}}$ and $\sum_{j=1}^{n} m_{\circ j} = m$. These observations are sufficient for obtaining estimates for marginal distributions of the individual components $Y_1, Y_2, \ldots, Y_n$. A real-life example for this case corresponds to frequency of long-distance phone calls m_{ij} placed at a certain geographical location Y_i at some given time unit y_{ij}.

Consider the element partition $\boldsymbol{p}$ $\{S_1, S_2, \ldots, S_n\}$ of S where $S_j = \{Y_j\}$ for each component Y_i. Thus, the marginal probability of each component $S_j = \{Y_j\}$ is $P(S_j) = \pi_j$, which can be consistently estimated by $\widehat{\pi}_j = \frac{m_{\circ j}}{m}$. For each component Y_i,, the probability of the event $\{Y_j = y_{ij}\}$ is represented by $P(Y_j = y_{ij} | S_j) = \pi_{i|j}$, which will, similarly be estimated by $\widehat{\pi}_{i|j} = \frac{m_{ij}}{m_{\circ j}}$. However, after observing the system S, the probability for the same event $\{Y_j = y_{ij}\}$ becomes $P(Y_i = y_{ij} | S) = \pi_{ij}$ and is estimated with $\widehat{\pi}_{ij} = \frac{m_{ij}}{m}$. Accordingly, estimates for the entropies of individual components are

$$\boldsymbol{H}(\widehat{\boldsymbol{S}_j}) = -\sum_{\boldsymbol{i}=1}^{\kappa} \frac{\boldsymbol{m_{ij}}}{\boldsymbol{m_{\circ j}}} \log\left(\frac{\boldsymbol{m_{ij}}}{\boldsymbol{m_{\circ j}}}\right), \quad j = 1, 2, \ldots, \mathrm{n}$$

and the estimate for the system entropy is given by

$$H(S_1, \widehat{S_2}, \ldots, S_n) = -\sum_{j=1}^{n} \sum_{i=1}^{\kappa} \frac{m_{ij}}{m} \left((log \frac{m_{ij}}{m}) \right).$$

By (11), the estimated information transfer will be

$$Tr(\widehat{S},P)=\sum_{j=1}^{n}\boldsymbol{H}(\widehat{\boldsymbol{S}_j})-H(S_1,\widehat{S_2},\ldots,S_n)$$
$$=-\sum_{i=1}^{\kappa}\frac{\boldsymbol{m_{ij}}}{\boldsymbol{m_{\circ j}}}\log(\frac{\boldsymbol{m_{ij}}}{\boldsymbol{m_{\circ j}}})+\sum_{j=1}^{n}\sum_{i=1}^{\kappa}\frac{m_{ij}}{m}\left((log\frac{m_{ij}}{m})\right) \quad (17)$$

A vanishing value of this non-negative real number in (17) provides some evidence that the partition $\boldsymbol{p}\ \{S_1, S_2,\ldots,S_n\}$ conforms well with the given system *S*, so that the degree of its divergence from zero is a clue that there is still some unused information in $\boldsymbol{p}\ \{S_1, S_2,\ldots,S_n\}$ for *S*. When $Tr(\hat{S},P)= 0$, observations suggests that the system cannot be decomposed further than the element decomposition $\boldsymbol{p}\ \{S_1,S_2,\ldots,S_n\}$. A non-vanishing value of the empirical information transfer function $Tr(\hat{S},P)$thus suggests that it is worthwhile to seek decompositions other than $\boldsymbol{p}\ \{S_1,S_2,\ldots,S_n\}$. We can partition *S* in some other way such as $\boldsymbol{p}\ \{S_1, S_2,\ldots,S_q\}$ where $1< q < n$. The foregoing analysis can be repeated using (14) to check whether $\boldsymbol{p}\ \{S_1, S_2, \ldots,S_q\}$ has some information for *S*. In that case, the system entropy estimate $H(S_1,\widehat{S_2},\ldots,S_n)$ will stay put, but estimates of entropies $H(\hat{S}_\ell)$ of individual subsets S_ℓ will change.

The given observations in the Table 1 are clearly not sufficient for obtaining a conclusion beyond the result obtained above. For further conclusions we need more sophisticated observations and experiments. Hence, we have the next illustration:

Illustration Two

The second illustration relates to a case where empirical data for pairs of components are available. This case is more general in the sense that availability of empirical observations (frequencies) on pairs of random variables also implies availability of observations on individual variables. Assume again, for simplicity of exposition, that the random components $Y_1, Y_2,\cdots, Y_{n-1}$ and Y_n are discreet and the number of functional values taken by each Y_i is r_i, $(i=1,2,\ldots,n)$ with these values ranging over the whole numbers

$$y_{i1}, y_{i2},\ldots, y_{ir_i}$$

For notational convenience and without a loss of generality, we can assume that $r_i = r$ for all $i=1,2,\ldots,n$. In accordance with a certain pattern, the joint event $\{Y_i=y_{ih}, Y_j=y_{jk}\}$, $i\bullet j$, takes values in the discrete $(r(n-1))\times(r(n-1))$ matrix layout set up as in Table 2. This layout is now our observational system *T*, which is a proper subset of $S\times S$.

Because we are interested in cross pairs of variables like Y_i and Y_j with $i \neq j$, the diagonal cells (darker gray) are irrelevant and we therefore have (n−1) rows and (n−1) columns of interest. Hence, off-diagonal cells become a center of interest. Each off-diagonal cell is composed of an $(r\times r)$ matrix of observations on the bi-variable event such as $\{Y_i=y_{ih}, Y_j=y_{jk}\}$, $i \neq j = 1,2,\ldots,n$ and $h,k = 1,2,\ldots,r$. When the two events $\{Y_i=y_{ih}, Y_j=y_{jk}\}$ and $\{Y_j=y_{jk}, Y_i=y_{ih}\}$ are identical, i.e., when they are symmetrical, only the upper or lower off-diagonal part of the table can be used. For convenience of visualization, the unused part (for instance, the lower part) is shaded in light gray. In this latter case, the upper off-diagonal block cells become the system *T* of observations to be considered below. The bottom row and the last column shaded in blue are empty and the cell in the southeast corner contains one single element marked $\boldsymbol{m}$ representing total number of observations.

To aid exposition, the cell corresponding to the observational frequencies on the i^{th} and j^{th} variables (orange) is reproduced in Table 3 with the same color. For the specific pair of Y_i and Y_j, the symbol $m_{i(h)j(k)}$ in stands for the frequency with which the bi-variable event $\{Y_i=y_{ih}, Y_j=y_{jk}\}$ is observed empirically. The last column and the bottom row present marginal frequencies involved (column-wise and row-wise sums of the frequencies in the table): For observations on

$\{Y_i=y_{is}\}$ corresponding to a certain $s = 1,2,\ldots,r$ we have

$$m_{i(s)j}=\sum_{k=1}^{r} m_{i(s)j(k)},\quad i \neq j$$

and similarly, for a certain $t = 1,2,\ldots,r$ we have

$$m_{i\,j(t)}=\sum_{h=1}^{r} m_{i(h)j(t)},\quad i \neq j$$

Thus,

$$m_{ij}=\sum_{i=1}^{r} m_{i(s)j}=\sum_{i=1}^{r} m_{i\,j(t)}$$

with

$$m=\sum_{i=1}\sum_{i<j} m_{ij}.$$

As it is designated in Table 2, m is located in the southeast cell of the table.

The $q=\frac{n(n-1)}{2}$ upper off-diagonal block cells of Table 2 can now be labeled $T_1, T_2, \ldots, T_q$. $\boldsymbol{p}$ $\{T_1, T_2, \ldots, T_q\}$ is a proper partition of T. Labeling starts from the top leftmost cell of the table and then goes to the leftmost off-diagonal block cell of the second row and so on as in the lexicographical ordering: T_1 for $\{Y_1,Y_2\}$, T_2 for $\{Y_1,Y_3\},\ldots,$ T_q for $\{Y_{n-1},Y_n\}$. For each partition $\ell = (i,j)$, $i < j$, estimates of the marginal probabilities

$$P(Y_i=y_{ih},\ Y_j=y_{jk})=\pi_{i(h)j(k)},\quad j>i=1,2,\ldots,\mathrm{n}\ \text{and}\ h,k=1,2,\ldots,\mathrm{n}$$

are given by

$$\widehat{\pi}_{i(h)j(k)|\ell}=\frac{m_{i(h)j(k)}}{m_{ij}}$$

For the whole system the same estimate becomes

$$\widehat{\pi}_{i(h)j(k)}=\frac{m_{i(h)j(k)}}{m},$$

so that, as before,

$$H(\widehat{T}_\ell))=-\sum_{h=1}^{r}\sum_{k=1}^{r}\frac{m_{i(h)j(k)}}{m_{ij}}log(\frac{m_{i(h)j(k)}}{m_{ij}})$$
$$H(T_1,\widehat{T}_2,\ldots,T_q)=-\sum_{i=1}^{n}\sum_{i<j}^{n}\sum_{h=1}^{r}\sum_{k=1}^{r}\frac{m_{i(h)j(k)}}{m}log(\frac{m_{i(h)j(k)}}{m})$$

The estimated transfer function

$$\widehat{T}r(T,P)=\sum_{\ell=1}^{q}H(\widehat{T}_\ell)-H(T_1,\widehat{T}_2,\ldots,T_q)$$

will have small values close to zero when a pair-wise partition is enough for decomposition whereas its higher values will produce evidence that there is further information to be utilized in T for a further partition $\boldsymbol{p}$ $\{T_1, T_2, \ldots, T_r\}$, $q \neq r$.

ITF seems to be free from restrictions of structural modeling and does not depend on the type of distribution of the random variables involved. It also does not require any restriction on the type of interaction of variables such as their uncorrelatedness, independence, etc. It therefore is a flexible technique but need a great deal of care and designing efforts before being applied empirically. As it is the case with previous approaches, high dimensionality of data may at times be a deterrent factor in its application.

Concluding Remarks

This work is an attempt to provide a uniform body of exposition for available techniques of decomposition in the literature. Because of prevailing high-dimensionality issues, these techniques have increased in importance for the past two decades. Except for some comments and the two illustrations of the use of information transfer and for our efforts to present a unified standpoint, we do not claim originality. Our unifying framework has been probability theory and multivariate statistics as well as information theory. We have not dwelt on inferential issues concerning use of statistical estimates because of our intension to provide a general text that will apply to a large spectrum of application areas.

Regarding the applicability of the five techniques reviewed: All of the four techniques PCA, SVD, ICA, and NCA are based on structural modeling of a real-life phenomenon; they are dependent upon some assumptions concerning interactive nature of real-life events such as uncorrelatedness or independence which may be restrictive for data; they are analytically well-founded, which paves the way for their preference; and they have dimension reduction properties in their favor. The last technique, ITF, is free from all such drawbacks but requires challenging tasks of design and computation. Unlike the first three techniques, it is not particularly used for data reduction purposes. All five techniques confront high-dimensionality problems and seem to need care in applications involving Gaussian distributions.

Acknowledgments

We would like to acknowledge and are thankful for some constructive comments and suggestions of two anonymous referees with whose contributions the manuscript was substantially improved. The first two authors would also like to express their gratitude to the Department of Electrical and Computer Engineering of University of Alabama at Birmingham for providing a convenient administrational setting for cooperation.

References

Anderson, TW. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons; New York: 1984.

Arnold, B.; Sarabia, J-M.; Castillo, E. Conditional Specification of Statistical Models: Models and Applications, Springer Series in statistics. Springer-Verlag; New York: 1999.

Ash, RB. Information Theory. Dover; New York: 1990.

Barlow, RE.; Prochan, F. Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston; New York: 1975.

Browning TR. Applying the design structure matrix to system decomposition and integration problems: A review and new directions. IEEE Transactions on Engineering Management 2001;48 (3):292–306.

Butterfly P, Sudbery A, Szule J. Compatibility of subsystem states. Quantum Physics 2005;3 (arXiv:quant-ph/0407227 v 3 22 Apr 2005)

Chung, S.; Seabrook, C. A singular value decomposition: analysis of grade distributions. Georgia Institute of technology; VIGRE REU: 2004. http://www.its.caltech.edu/~mason/research/carstep.pdf

Comon P. Independent component analysis, a new concept? Signal Precessing 1994;36:287–314.

Conant, RG. Information transfer in complex systems with applications to regulation. University of Illinois; 1968. unpublished Ph.D. dissertation

Conant RG. Detecting subsystems of a complex system. IEEE Transactions on Systems, Man, and Cybernetics 1972 September;:550–553.

Dall'Aglio. Frechet classes and compatibility of distribution functions. Symposia Math 1972;9:131–150.

Donoho, S. High-dimensional data analysis: the blessings and curses of dimensionality. AMS special conference "0n mathematical challenges of the 21st century"; UCLA, Los Angeles. 6–11 August 2000; 2000.

Eckart C, Young G. Approximation of one matrix by another of lower rank. Psychometrika 1936;1:211–218.

Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. The Mathematics Arxive: math. 2006ST/0602133

Gurrera, MDC. Construction of bivariate distributions and statistical dependence operations. University of Barcelona Department of Statistics; 2005. Ph. D. Dissertation

Hastie T, Tibshirani R, Eisen ME, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology 2000:1. [PubMed: 11178226]

Hotelling H. Analysis of a complex of statistical variables with principal components. Journal of Educational Psychology 1933;24:417–441. 498–520.

Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. Neural Networks 2000;13:411–430. [PubMed: 10946390]

Hyvarinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley and Sons; New York: 2001.

Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. Vol. 5. Prentice-Hall, Inc; Englewood Cliff, New Jersey: 2002.

Jolliffe, IT. Principal Component Analysis. Vol. 2. Springer-Verlag; New York: 2002.

Jutten C, Herault J. Blind separation of sources, Part 1: An adaptive algorithm based on neuromimatic architecture. Signal Processing 1991;24:1–10.

Kendall, M. Multivariate Analysis. Charles Griffin & Company; London: 1975.

Kullback, S. Information and Statistics. John Wiley and Sons; New York: 1959.

Lawson, CL.; Hanson, RJ. SIAM: Classics in Applied Mathematics. Vol. 3. Vol. 15. Philadelphia: 1995. Solving Least Squares Problems.

Lee TW, Girolami M, Bell AJ, Sejnowski TJ. A unifying information theoretic framework for independent component analysis. International Journal on Mathematical and Computer Modeling. 1998

Liao R, Boscolo Y-L, Yang LM, Tran CS, Roychowdhury VP. Network component analysis. PNAS 2003;100:15522–15527. [PubMed: 14673099]

Papoulis, A.; Pillai, SU. Probability, Random Variables and Stochastic Processes. Vol. 4. McGraw-Hill; 2002.

Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 1901;2:559–572.

Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical Recipes in C. Vol. 2. Cambridge University Press; Cambridge: 1992.

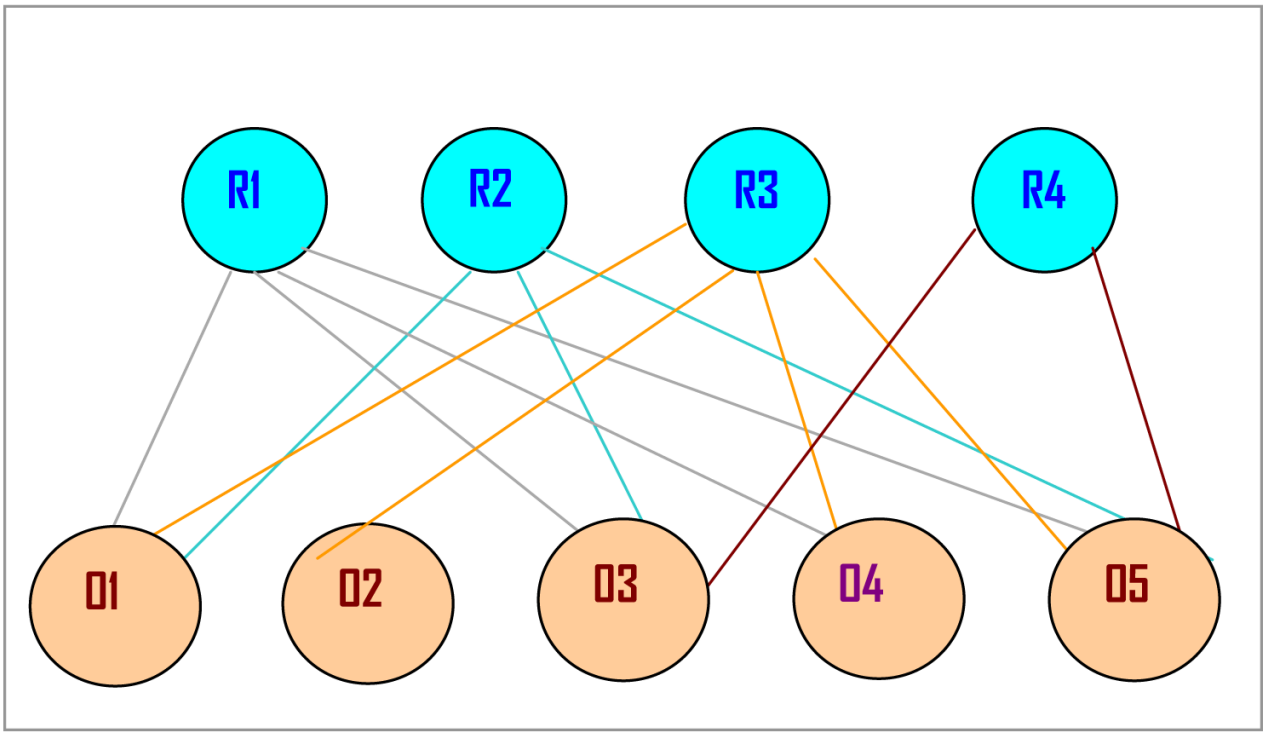
Scheffe, H. The Analysis of Variance. John Wiley and Sons; New York: 1959.

Schroedinger, E. What Is Life ?. Cambridge University Press; Cambridge: 1944.

Seber, GAF. Multivariate Observations. John Wiley and Sons; New York: 1984.

Srinivasan, SH.; Ramakrishnan, KR.; Budhlakoti, S. Character decompostions. www.ee.iitb.ac.in/~icvgip/PAPERS/292.pdf

Stewart GW. On the early history of the singular value decomposition. SIAM Review 1993;35(4):551–566.



	R1	R2	R3	R4
O1	1	1	1	0
O2	0	0	1	0
O3	1	1	0	1
O4	1	0	1	0
O5	1	1	1	1

FIGURE 1. A Pictorial Example for Network Design

Depicted in the upper part is a bipartite network between four regulatory agents (gene knock-outs, drugs etc.) denoted by **R1** through **R4** and five outputs (e.g., growths of certain organisms) designated with O1 through **O5**. Connections between regulatory nodes and output nodes are shown as usual. The rectangular table just underneath the network represents the connectivity matrix **B** with columns corresponding to regulatory agents and rows to outputs. The regulatory matrix **C** is not shown in the picture. When there are no connections between output nodes (rows) and regulatory nodes (columns), the corresponding cells contain zero. Obviously, all other cells are composed of ones for the network shown here.

Table 1

Observations on element partitioning

	Y_1	Y_2	…	Y_n	
y_{i1}	m_{11}	m_{12}	…	m_{1n}	
y_{i2}	m_{21}	m_{22}	…	m_{2n}	
⋮	⋮	⋮	⋮	⋮	
y_{ik}	m_{k1}	m_{k2}	…	m_{kn}	
Marginal Sums	$m_{\cdot 1}$	$m_{\cdot 2}$	…	$m_{\cdot n}$	m

Table 2

Pair-wise observations

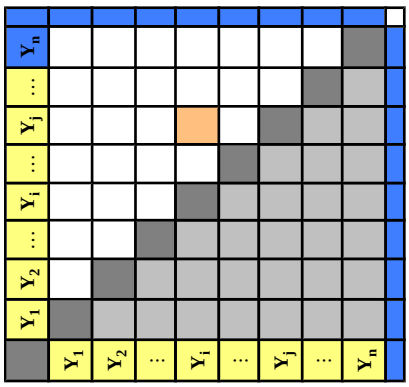


Table 3

Frequencies of $\{Y_i=y_{ih}, Y_j=y_{jk}\}$ in Table 2 for the highlighted pair of i and j

$m_{i(1)j(1)}$	$m_{i(1)j(2)}$	…	$m_{i(1)j(r)}$	$m_{i(1)j}$
$m_{i(2)j(1)}$)	$m_{i(2)j(2)}$	…	$m_{i(2)j(r)}$	$m_{i(2)\,j}$
⋮	⋮	⋮	⋮	⋮
$m_{i(r)j(1)}$	$m_{i(r)j(2)}$	…	$m_{i(r)j(r)}$	$m_{i(r)\,j}$
$m_{i\,j(1)}$	$m_{i\,j(2)}$	…	$m_{i\,j(r)}$	$m_{i\,j}$