withdrawing substituents. The measurement of lipophilicity, using experimental partition coefficient data, was more dependable than an estimation from literature data collected in a different system. The preference for actual measurement over the estimation of lipophilicity is even more obvious when one deals with drug molecules having a complicated molecular structure.

REFERENCES

(1) C. Hansch, *Accounts Chem. Res.*, **2**, 232(1969).

(2) W. P. Purcell, J. A. Singer, K. Sundarm, and G. L. Parks, in "Medicinal Chemistry," A. Burger, Ed., Wiley, New York, N.Y., 1970, p. 164.

(3) T. K. Lin, *J. Med. Chem.*, **17**, 151(1974).

(4) C. Hansch and W. J. Dunn, III, *J. Pharm. Sci.*, **61**, 1(1972).

(5) A. Goldstein, *Pharmacol. Rev.*, **1**, 102(1949).

(6) C. Hansch, K. Kiehs, and G. Lawrence, *J. Amer. Chem. Soc.*, **87**, 5770(1965).

(7) J. M. Vandenbelt, C. Hansch, and C. Church, *J. Med. Chem.*, **15**, 787(1973).

(8) A. E. Bird and A. C. Marshall, *Biochem. Pharmacol.*, **16**, 2275(1967).

(9) F. Helmer, K. Kiehs, and C. Hansch, *Biochemistry*, **7**, 2858(1968).

(10) T. Fujita, *J. Med. Chem.*, **15**, 1049(1972).

(11) T. K. Lin, Y. W. Chien, R. R. Dean, J. E. Dutt, H. W. Sause, C. H. Yen, and P. K. Yonan, *ibid.*, **17**, 751(1974).

(12) H. W. Sause, J. W. Cusic, and P. K. Yonan, "Anti-Arrhythmic Activity of 4-Dipropylamino-2-aryl-2-pyridyl Butyramides," presented at 165th National Meeting of American Chemical Society, Dallas, Tex., Apr. 1973.

(13) C. H. Yen, H. S. Lowrie, and R. R. Dean, *J. Med. Chem.*, **17**, 1131(1974).

(14) J. W. Panter, D. W. Boykin, Jr., and D. Wilson, *ibid.*, **16**, 1366(1973).

(15) J. J. Fischer and O. Jardetzky, *J. Amer. Chem. Soc.*, **87**, 3237(1965).

(16) J. G. Kirkwood and I. Oppenheim, "Chemical Thermodynamics," McGraw-Hill, New York, N.Y., 1961, chaps. 8, 11.

(17) J. T. Edsall and J. Wyman, "Biophysical Chemistry," vol. 1, Academic, New York, N.Y., 1958, p. 160.

(18) T. Fujita, J. Iwasa, and C. Hansch, *J. Amer. Chem. Soc.*, **86**, 5175(1964).

ACKNOWLEDGMENTS AND ADDRESSES

Received June 20, 1974, from the *Biopharmaceutics Section, Development Department, Searle Laboratories, Division of G. D. Searle & Company, Skokie, IL 60076*

Accepted for publication November 21, 1974.

Appreciation is expressed to Dr. H. W. Sause, Mr. J. H. Yen, Mr. P. K. Yonan, and Dr. J. W. Cusic for donating disopyramide derivatives, to Miss D. M. Jefferson for technical assistance, and to Mrs. M. Fisher for manuscript preparation.

[x] To whom inquiries should be directed.

Selecting Key Parameters in Pharmaceutical Formulations by Principal Component Analysis

N. R. BOHIDAR [x], F. A. RESTAINO, and J. B. SCHWARTZ

Abstract □ The role of principal component analysis in the selection of pharmaceutical formulations is presented. The objective and the procedure of the analysis are discussed in detail. The technique was successfully applied to a system consisting of 10 response variables (tablet properties). Analysis of the results showed that the first component (dissolution) and components one and two together (dissolution and disintegration) contributed 95.4 and 99.3%, respectively, to the overall information about the formulations and that eight of 10 response parameters contributed nothing further to the overall information. The results obtained from this method of analysis may be found useful for achieving economy in both cost and time of measuring responses. Principal component analysis also provides a basis for understanding the underlying mechanism of the system under consideration.

Keyphrases □ Pharmaceutical formulations—selecting key parameters by principal component analysis, examples □ Formulations, pharmaceutical—selecting key parameters by principal component analysis, examples □ Principal component analysis—selecting key parameters in pharmaceutical formulations

In the development of a drug delivery system, a research pharmacist usually measures several response parameters. For instance, 10 or more parameters were considered in the development of a pharmaceutical tablet formulation (1). Based on all of these parameters, one attempts to find those levels of the formulation factors (diluent ratio, compressional force, *etc.*) for which the system is considered optimum. Since a large number of interrelated response variables is generally involved, it is relevant to ask how the interrelation and covariation of these measurements might be represented and whether fewer measurements might not carry all the necessary information for accomplishing a specific objective.

When several formulations are available, the developmental pharmacist must determine how best to distinguish between them. When one is choosing between two or three, the trend may be obvious. For example, the formula changes made may cause no difference in tablet hardness but considerable difference in disintegration characteristics. But when a long list of formulations is available or, more precisely, when one has infinite possibilities (as in computer optimization) and is dealing with many parameters, the trend is less obvious.

One may have certain basic constraints, such as a minimum hardness value, but it is nevertheless important to know which property or properties can be used to distinguish between choices. Generally, an educated guess is made, based on experience with the system and with pharmaceutical systems in general.

But there is a mathematical method to select those variables that best distinguish between formulations and those variables that change most drastically from

one formulation to another and which should be the criteria upon which one selects a formulation.

A multivariate statistical procedure called "principal component analysis" can effectively be used to answer these questions. Basically, this method first finds the function of the observations that has the largest variance and then finds the function, independent of the first function, having the largest variance. It then finds the function, independent of the first two functions, with the largest variance and so on.

If p variates (responses) are observed, it is obvious that p of these could account for all of the variability in the original observations. The problem is whether fewer variates might be used. If so, some principal components might not contribute anything to the overall variability; thus, they would not help in distinguishing one formulation from another.

Large sets of data should not be subjected to a principal component analysis merely to obtain fewer variables to work with without regard to an overall objective. The objective should be established first and then the principal component model should be used only if it complements the objective. However, tremendous simplification of the data can often be effected for problems where principal component analysis is appropriate.

The primary purpose of this paper is to demonstrate the role of principal component analysis in pharmaceutical formulation development by applying the procedure to the data from an optimization experiment previously described (1). Ten response variables were measured on each of 27 formulations, and the variables substantially contributing to the overall information were identified.

The theoretical section is meant to serve as a guide for following the steps in the principal component analysis technique. However, computer programs can perform all of the indicated operations, and it is only necessary to feed in the raw data. A familiarity with the theory is useful for analysis of the results and for those who wish to modify available programs.

EXPERIMENTAL

The 10 parameters shown in Table I were measured on each of the 27 tablet formulations considered for the experiment[1] described in detail in Ref. 1. Thus, the data set to be subjected to principal component analysis contains 27 values for each response shown in Table I.

THEORY AND PROCEDURE

The structure of the data associated with the experiment is presented in the matrix form shown in Table II.

For the purpose of describing the procedure, let Y_{ik} ($i = 1, 2, \ldots, p$; $k = 1, 2, \ldots, N$) denote the numerical value associated with the kth experiment for the ith response variable, and let p and N represent the number of response variables and the number of experiments for each parameter considered, respectively. In this case, $p = 10$ and $N = 27$ in Table II. The value of Y_{ik} can be the mean of several measurements.

[1] The technique utilized here may be applied to any multivariate system and need not be limited to the statistically designed set of experiments considered here.

Table I—Response Variables

Symbolic Designation	Response Variable	Units
Y_1 (DT)	Disintegration time	Minutes
Y_2 (HD)	Tablet breaking strength	Kilograms
Y_3 (DR)	Dissolution	Percent released in 30 min
Y_4 (FR)	Friability	Percent weight loss
Y_5 (TH)	Thickness uniformity	*RSD*, %
Y_6 (PO)	Porosity	Milliliters per gram
Y_7 (MP)	Mean pore diameter	Micrometers
Y_8 (WT)	Weight uniformity	*RSD*, %
Y_9 (TB)	Tablet breakage	Number of chipped tablets
Y_{10} (GM)	Granulation mean diameter	Millimeters

The variance (S_{ii}) and the covariances (S_{ij}, $i \neq j$) associated with the p parameters are then calculated as follows:

$$\text{variance} = S_{ii} = \left[\sum_{k=1}^{N} Y_{ik}^2 - \left(\sum_{k=1}^{N} Y_{ik}\right)^2 N^{-1}\right](N-1)^{-1} \quad \text{(Eq. 1)}$$

where $i = 1, 2, \ldots, p$, and:

$$\text{covariance} = S_{ij}(i \neq j) = \left[\sum_{k=1}^{N} Y_{ik}Y_{jk} - \left(\sum_{k=1}^{N} Y_{ik}\sum_{k=1}^{N} Y_{jk}\right)N^{-1}\right](N-1)^{-1} \quad \text{(Eq. 2)}$$

$$= S_{ji}$$

where $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$.

There would be a total of p variances and $\frac{1}{2}p(p-1)$ covariances associated with p parameters. There will be 10 variances and 45 covariances when 10 parameters are considered. The variance and covariance quantities are then arranged in a square matrix form as follows:

$$\Sigma = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1p} \\ S_{21} & S_{22} & \ldots & S_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ S_{p1} & S_{p2} & \ldots & S_{pp} \end{bmatrix} \quad \text{(Eq. 3)}$$

This matrix is called the variance–covariance matrix and is denoted by the greek letter Σ (sigma). The variances are arranged in the main diagonal of the matrix and the covariances are placed in their respective off-diagonal positions. Since $S_{ij} = S_{ji}$ ($i \neq j$), one has only $\frac{1}{2}p(p+1)$ distinct elements in the matrix. If there are p parameters, then the dimension of this matrix is $(p \times p)$ with p rows and p columns and there are p^2 elements in the matrix. When $p = 10$, the dimension of Σ is (10×10) with 100 elements in the matrix.

The determinant of the matrix reduces these p^2 (here 100) elements to a single number. This number represents the variance of the entire system, usually called the generalized variance. This statistic is helpful for comparing the variances of two different systems. The primary interest here, however, is in the magnitude of the variances of the individual components and their relative information within the system under consideration. So consider the determinant of the matrix $(\Sigma - \lambda I)$, where I is the identity matrix with ones in the main diagonal and zeros elsewhere and, expressed explicitly, gives:

$$|\Sigma - \lambda I| = \begin{vmatrix} S_{11} - \lambda_1 & S_{12} & \ldots & S_{1p} \\ S_{21} & S_{22} - \lambda_2 & \ldots & S_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ S_{p1} & S_{p2} & \ldots & S_{pp} - \lambda_p \end{vmatrix} \quad \text{(Eq. 4)}$$

Table II—Data Matrix

Experiment Number (Formulation)	Response Variables									
	(DT)	(HD)	(DR)	(FR)	(TH)	(PO)	(MP)	(WT)	(TB)	(GM)
1	Y_{11}	Y_{21}	Y_{31}	Y_{41}	Y_{51}	Y_{61}	Y_{71}	Y_{81}	Y_{91}	$Y_{10,1}$
2	Y_{12}	Y_{22}	Y_{32}	Y_{42}	Y_{52}	Y_{62}	Y_{72}	Y_{82}	Y_{92}	$Y_{10,2}$
3	Y_{13}	Y_{23}	Y_{33}	Y_{43}	Y_{53}	Y_{63}	Y_{73}	Y_{83}	Y_{93}	$Y_{10,3}$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
27	$Y_{1,27}$	$Y_{2,27}$	$Y_{3,27}$	$Y_{4,27}$	$Y_{5,27}$	$Y_{6,27}$	$Y_{7,27}$	$Y_{8,27}$	$Y_{9,27}$	$Y_{10,27}$

In this equation, the λ_i's ($i = 1, 2, \ldots, p$) are unknown. By setting:

$$|\Sigma - \lambda I| = 0 \qquad \text{(Eq. 5)}$$

and expanding the determinant on the left, one has a polynomial equation of pth order of the following form:

$$|\Sigma - \lambda I| = f(\lambda) = (-\lambda)^p + a_{p-1}(-\lambda)^{p-1} + \ldots + a_1(-\lambda) + a_0 = 0 \qquad \text{(Eq. 6)}$$

where the coefficients a_i's are, but for signs, the sum of all of the principal ith-order minors of the determinant of the variance–covariance matrix Σ. A pth-order polynomial would yield p roots (zeros of the polynomial). These roots (λ_i's) are known as characteristic roots, latent roots, or eigenvalues. The term "eigenvalue" will be used in the subsequent reference to the roots. These eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_p$) represent the variances of each "orthogonal" component of the system. The variance–covariance matrix of the orthogonal system (Λ) has the following structure:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \ldots & \lambda_p \end{bmatrix} \qquad \text{(Eq. 7)}$$

It is clearly seen in Eq. 7 that the original system has been transformed to an orthogonal system in which the covariances of the components have been reduced to zero (hence the term orthogonal system is used). The generalized variance of the original system is identically equal to the generalized variance of the orthogonal system, that is:

$$|\Sigma| = |\Lambda| \qquad \text{(Eq. 8)}$$

This clearly indicates the complete preservation of the total information of the original system in the transformed system. The orthogonal system provides an estimation of the relative contribution of each component to the overall information. Let $\theta = \lambda_1 + \lambda_2 + \ldots + \lambda_p$; then $100\lambda_1\theta^{-1}, 100\lambda_2\theta^{-1}, \ldots, 100\lambda_p\theta^{-1}$ are the respective relative contribution (in percent) of each of the p components of the system to the overall information.

Now consider the structure of each component. Associated with each eigenvalue λ_r ($r = 1, 2, \ldots, p$) is a set of p coefficients, $\nu_{r1}, \nu_{r2}, \ldots, \nu_{rp}$; $r = 1, 2, \ldots, p$. This set of coefficients, called the eigenvector of eigenvalue λ_r, is a solution of the equation:

$$(\Sigma - \lambda_r I)\nu = 0 \qquad r = 1, 2, \ldots, p \qquad \text{(Eq. 9)}$$

where I is the identity matrix with ones in the main diagonal and zeros elsewhere. This equation is expressed explicitly as follows:

$$\begin{bmatrix} S_{11} - \lambda_r & S_{12} & \ldots & S_{1p} \\ S_{21} & S_{22} - \lambda_r & \ldots & S_{2p} \\ . & & & \\ . & & & \\ . & & & \\ S_{p1} & S_{p2} & \ldots & S_{pp} - \lambda_r \end{bmatrix} \begin{bmatrix} \nu_{r1} \\ \nu_{r2} \\ . \\ . \\ . \\ \nu_{rp} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} \qquad \text{(Eq. 10)}$$

where $r = 1, 2, 3, \ldots, p$.

The magnitudes of the coefficients associated with those principal components that are substantially contributing to the overall information are of interest in the interpretation of the principal component analysis of the system under consideration.

The codes of a FORTRAN program for computing the eigenvalues and eigenvectors of a real symmetric matrix are available (2) and are suitable for adoption into any computer system without considerable effort.

The detailed mathematics of the procedure can be found in Ref. 3.

RESULTS AND DISCUSSION

The results of the principal component analysis associated with the system under consideration are presented in Tables III–V. The results presented in Tables III and IV pertain to the situation in which all 10 parameters were included in the analysis, as does Analysis I in Table V. The other results presented in Table V pertain to situations where two or more parameters were excluded from the system. This approach is not an essential feature of the principal component analysis. However, it is provided here to illustrate the results of the analysis in the presence or absence of cer-

Table III—Eigenvalues, Relative Information, and Cumulative Relative Information Associated with Each Component

Principal Component	Eigenvalues (λ_i)	Relative Information, % ($100\lambda_i\theta^{-1}$)	Cumulative Relative Information, %
I	578.4	95.4	95.4
II	23.9	3.9	99.3
III	2.2	0.4	99.7
IV	2.0	0.3	100.0
Total		100.0	

Table IV—Coefficients of Eigenvectors Associated with First Two Principal Components

Parameter	Principal Component I	Principal Component II
Dissolution (DR)	0.98	0.22
Porosity (PO)	0.00	0.00
Friability (FR)	0.00	0.00
Disintegration time (DT)	−0.22	0.97
Weight uniformity (WT)	−0.01	0.03
Thickness uniformity (TH)	0.00	0.00
Granular mean diameter (GM)	0.00	0.00
Tablet breakage (TB)	0.00	0.00
Mean pore diameter (MP)	0.00	0.00
Tablet breaking strength (HD)	0.01	0.12
Relative information, %	95.4	3.9

Table V—Relative Information (Percent) of Parameters for Five Principal Component Analyses

Serial Number	Analysis I		Analysis II		Analysis III		Analysis IV		Analysis V	
	Order of Parameters	Relative Information	Order of Parameters	Relative Information	Order of Parameters	Relative Information	Order of Parameters	Relative Information	Order of Parameters	Relative Information
1	(DR)	95.4	(DR)	95.4	—	—	—	—	—	—
2	(DT)	3.9	(DT)	3.9	—	—	—	—	—	—
3	(WT)	0.4	(WT)	0.4	(HD)	52.5	—	—	—	—
4	(HD)	0.3	(HD)	0.3	(WT)	44.9	—	—	—	—
5	(FR)	0.0[a]	(FR)	0.0	(FR)	2.3	(FR)	82.9	—	—
6	(MP)	0.0	—	—	(MP)	0.2	(MP)	10.0	(TB)	61.0
7	(TH)	0.0	—	—	(TH)	0.1	(TH)	4.5	(MP)	22.0
8	(TB)	0.0	—	—	(TB)	0.0	(TB)	1.4	(TH)	13.1
9	(GM)	0.0	—	—	(GM)	0.0	(GM)	1.1	(GM)	3.7
10	(PO)	0.0	—	—	(PO)	0.0	(PO)	0.1	(PO)	0.2
	Total	100.0		100.0		100.0		100.0		100.0

[a] Represents a rounded figure.

tain parameters. The information-analysis (variance-analysis) of the first four components is given in Table III. The structure-analysis of the first two components is presented in Table IV.

Table III lists the first four components, their respective eigenvalues, and the relative information calculated from them. An examination of these data shows that the total information contained in the system was contributed by these first four of the 10 principal components. In other words, the last six principal components did not contribute anything to the overall information. Furthermore, the first principal component contributed as much as 95.4% of the total information.

It is necessary now to conduct the structure-analysis of the components. The results presented in Table IV for principal component I reveal that "dissolution" was the predominant parameter of the component by sharing the largest value (0.98) among the coefficients associated with the component. Now, by relating the results of the variance-analysis and the structure-analysis, it may be inferred that dissolution accounted for most variabilities in the system. In other words, dissolution contained the most information concerning the power of distinguishing among the 27 formulations under consideration. If the objective is to achieve the optimum levels of the formulation factors (1), then constraining the dissolution parameter would lead to a faster selection of the optimum formulation than would constraining any other parameter considered.

Component II in Table III contributed only 3.9% of the total information. A structural analysis (Table IV) reveals that "disintegration time" was the predominant parameter associated with this principal component. The two parameters, dissolution and disintegration, yielded a cumulative relative information of 99.3%. The same procedure can be followed for each component.

In Table V, the parameters are presented in the order of the relative information contributed. The results of the principal component analysis involving all parameters in the system are shown in Analysis I. A discussion of these results was already presented. Analysis II shows the results of the principal component analysis involving only the first five parameters. The results of Analyses I and II are identical with respect to the order and magnitude of the relative importance of the first five parameters.

The results of the principal component analysis involving only the last eight parameters are presented in Analysis III; weight uniformity and tablet breaking strength become the predominant parameters in this particular system. The results of the principal component analysis involving only the last six parameters are presented in Analysis IV; friability, mean pore diameter, and tablet thickness become the predominant parameters in this system. The results of the principal component analysis involving only the last five parameters are presented in Analysis V; mean pore diameter, tablet breakage, and tablet thickness are the predominant parameters of this particular system. These types of analyses would provide a basis for a general understanding of the underlying mechanisms of the particular system considered.

A word of caution is in order. The results and the conclusions associated with a principal component analysis pertain only to the specific system to which the analysis was applied. Although the procedure may be applied to any multivariate system, any extrapolation of the results of one system to another is not appropriate. Each system must be analyzed separately and the results so obtained must be interpreted independently.

CONCLUSION

In this study it was clearly shown that principal component analysis can play an important role in pharmaceutical formulation by identifying parameters that are substantially contributing to the overall information associated with the system. It was observed that the first principal component contributed as much as 95.4% of the total information. The first two principal components together contributed as much as 99.3% of the overall information. Dissolution was identified as the predominant parameter of the system. Therefore, this parameter alone could effectively be used in the comparison of candidate formulations and in the constraining operation (1) for a faster selection of the optimum formulation.

Disintegration was identified as the next important parameter. All other parameters did not contribute substantially to the overall information. This information is vital in that one would be able to achieve economy in cost and time of measuring responses. Principal component analysis could also be used to provide a basis for understanding the underlying mechanism of the system.

REFERENCES

(1) J. B. Schwartz, J. R. Flamholz, and R. H. Press, *J. Pharm. Sci.*, **62**, 1165(1973).

(2) B. Carnahan, H. A. Luther, and J. O. Wilkes, "Applied Numerical Methods," Wiley, New York, N.Y., 1969, pp. 255–259.

(3) T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," Wiley, New York, N.Y., 1958, pp. 272–287.

ACKNOWLEDGMENTS AND ADDRESSES

Received February 15, 1974, from the *Department of Pharmaceutical Research and Development and Biometrics Research, Merck Sharp & Dohme Research Laboratories, West Point, PA 19486*

Accepted for publication November 19, 1974.

The authors thank Mr. J. E. Allegretti and Dr. J. L. Ciminera for their interest in the application of multivariate statistical methods.

[x] To whom inquiries should be directed.