DETERMINING THE NUMBER OF INTERACTING SPECIES: SIGNIFICANT FACTOR ANALYSIS

Herbert R. HALVORSON

Division of Biochemical Research, Henry Ford Hospital, Detroit, MI 48202, U.S.A.

Received 21 January 1981

The technique of significant factor analysis is described. This technique yields statistically valid criteria for defining the minimum number of linearly independent entities in an array of data. Analyses of examples from the literature are compared with other methods for extracting this information. Analysis of the subunit dissociation of hemoglobin illustrates the extent and the limits to model-independent conclusions that can be drawn. The technique is readily applicable with programs available in standard statistical packages.

1. Introduction

A common problem concerns the enumeration of independent species in an interacting system. That is, given a two-dimensional array of experimental observations, what is the minimal number of linearly independent species required to describe the data within experimental error? This problem arises in a variety of contexts and, understandably, there have been several approaches to its solution.

The technique of matrix rank analysis was first applied by Wallace [1] and subsequently by Weber [2] and Ainsworth [3]. This method entails reducing the data matrix by a series of elementary transformations until all but r of the rows contain nothing but zeros. The rank of the matrix is r, which is also the number of linear independent species. Consideration of error propagation requires that the same series of elementary transformations be performed on a companion error matrix, whose elements are the estimated uncertainties of the elements of the data matrix. The rank of the data matrix is then the number of rows in which the elements are significantly greater than the corresponding elements of the transformed companion error matrix. Similar approaches have been devised entailing a systematic search for nonzero minors of the data matrix.

Alternatively, Magar [4,5] has suggested the use of principal component analysis or abstract factor analysis, established statistical techniques, to obtain the same information. A major advantage of this idea is that the necessary computer programming is already contained in the standard statistical packages available at most computer facilities, greatly reducing the demands on the experimentalist. A difficulty is the lack of a convenient and well-grounded test for the significance of the principal components or abstract factors extracted from the data matrix.

This communication describes an application of abstract factor analysis, retaining the advantages cited by Magar, which also provides a statistically valid test of the significance of each additional factor and an assessment of the residual error unaccounted for by the significant factors, independent of a priori knowledge about the uncertainties. The analysis is applied to ten examples from the literature, permitting comparison with other methods of determining the rank of a matrix. A set of previously unpublished data [6] has also been analyzed, providing an example of potential pitfalls in using the rank of the data

0301-4622/81/0000-0000/$02.50

matrix as a guide to the number of interacting species in subsequent detailed analyses.

2. Theory

2.1. A pictorial overview

The principle of factor analysis can be visualized by considering a data matrix which is $n \times 3$ ($n > 3$). For the sake of concreteness, this could represent absorbance measured at 3 wavelengths for n different solution conditions (pH, concentration, etc.). Using the numbers in each column as orthogonal coordinates, the data set can be represented as n points along a space curve in three-dimensional space. The first abstract factor is simply the best straight line through the space curve. Now project the deviations from this line onto a plane perpendicular to the line. The best straight line through the resulting two-dimensional distribution is the second abstract factor, and the third is found by projecting the remaining deviations onto a perpendicular line. It is important to note that this procedure merely produces a new orthogonal coordinate system and that the data are in no way altered by being represented in the new coordinate system. The concept is readily generalized to spaces of higher dimension.

2.2. Unweighted data

Consider an $n \times p$ array of experimental data, $\mathbf{A}$, where for convenience it is assumed that $n > p$. Formally, the problem concerns the equation

$$\mathbf{A} = \mathbf{C} \times \mathbf{B} + \mathbf{E}. \tag{1}$$

with no knowledge about the right-hand side of the equation beyond the assumption that $\mathbf{E}$, the error matrix, represents random error. Matrices $\mathbf{C}$ and $\mathbf{B}$ are $(n \times m)$ and $(m \times p)$ respectively, and the problem is to estimate m. Returning to the example of absorbance measurements at p wavelengths under n conditions, c_{ik} represents the concentration of absorbing species k at the ith set of conditions and b_{kj} denotes the extinction coefficient of species k at wavelength j. The total number of absorbing species is m. Anticipating that sufficient data have been gathered to overdetermine the system, then the rank of $\mathbf{C}$, $r_{\mathbf{C}} = \min(n, m) = m$ and $r_{\mathbf{B}} = \min(m, p) = m$. (This assignment and possible exceptions are treated more fully in section 4.) The rank of the product of these two matrices $r_{\mathbf{CB}} = \min(r_{\mathbf{C}}, r_{\mathbf{B}}) = m$. But $\mathbf{A}$ also includes $\mathbf{E}$, which has a rank $r_{\mathbf{E}} = p$ (invoking the assumption that $n \geqslant p$). The significant factor problem of estimating m in the presence of $\mathbf{E}$ can be seen as entirely analogous to the problem of determining the "best" degree for a polynomial fit to noisy data.

If $\mathbf{A}$ is premultiplied by its transpose $\mathbf{A}'$, the result, divided by n, is a $(p \times p)$ variance-covariance matrix describing the distribution about the origin

$$\mathbf{V}_{\mathrm{O}} = \mathbf{A}' \times \mathbf{A}/n. \tag{2}$$

Factor analysis now proceeds by solving the eigenvalue-eigenvector problem

$$\mathbf{V}_{\mathrm{O}}\mathbf{F} = \mathbf{F}\Lambda \quad (\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p). \tag{3}$$

The matrix Λ is a diagonal matrix of eigenvalues λ_j ("length" of the abstract factors) and the columns of $\mathbf{F}$ are the associated eigenvectors (orientations of the abstract factors and the transformed coordinate system).

In the absence of experimental error, all but m of the p eigenvalues will be zero. Conventional factor analysis of unweighted data has accordingly been concerned with assessing the number of eigenvalues which differ significantly from zero. This can be difficult and uncertain. The following procedure has been adopted instead. Each eigenvalue represents the amount of the variance associated with each abstract factor and the sum of the eigenvalues is the total variance. One seeks to partition the p eigenvalues into two sets: m eigenvalues associated with significant factors and $p - m$ eigenvalues associated with pure error. A significant factor is one which effects a significant reduction of the residual variance. Specifically, one forms the ratio of the improvement in the variance to the residual variance and applies the F-test. That is, to test the significance of the kth factor, one compares the value of

$$F_\alpha(1, p-k) = \frac{\lambda_k}{\Sigma_{k+1}^{p}\lambda_j/(p-k)}, \tag{4}$$

with the critical values tabulated in statistical tables. It is common statistical practice to use $F_{0.95}$ as the criterion for significance. This leaves a probability $P = 0.05$ that the result arises entirely from chance. If the measured value of F_α exceeds the tabulated value, then one accepts the factor as being significant. The rank $r_\mathbf{A} = m$ then becomes the largest value of k which passes the test for significance. Subsequently r_0 is used to denote the estimate of the rank derived from analysis of the distribution about the origin.

As an additional benefit to this procedure, one immediately obtains a best estimate of the residual uncertainty. The standard error is simply

$$\mathrm{SE} = \left(\sum_{m+1}^{p} \lambda_j / (p - m) \right)^{1/2}. \tag{5}$$

If this is unreasonably large (or small), it suggests that the estimate of m should be appropriately modified.

2.3. Weighted data

Frequently it is possible to estimate the uncertainty in the data (the elements of $\mathbf{E}$). In this case each element of $\mathbf{A}$ is replaced by

$$a_{ij} / \left[\mathrm{var}(a_{ij}) \right]^{1/2}$$

and the analysis proceeds as above. Here conventional factor analysis is concerned with detecting eigenvalues that exceed unity. Obviously, an inappropriate estimate of the uncertainty in the data can have disastrous effects. The use of eq. (4), however, is relatively immune to scaling errors, responding only to an inappropriate distribution of the estimated uncertainty.

This ancillary information permits the application of an additional statistical test related to eq. (5). The statistic $\Sigma_{m+1}^{p} \lambda_j / (p - m)$ can be tested as a reduced chi-square (χ_ν^2) with $p - m$ degrees of freedom to estimate the probability that the remaining variance arises from chance. If all elements of $\mathbf{E}$ have the same expectation and the estimated uncertainty has been determined with d degrees of freedom, then the more powerful F-test, $F_\alpha(p - m, d)$, can be applied to this statistic. In either event, one seeks a value of m such that the residual variance is not statistically significant. However, if the probability that the observation is due to chance approaches one, it invariably means that m has been overestimated. Obviously, this test is sensitive to the magnitude of the estimated uncertainty in the data.

2.4. Covariance about the mean

As an adjunct to the analysis described above, it is useful to perform factor analysis on the variance about the mean, $\mathbf{V}_\mathrm{M}$ as well:

$$\mathbf{V}_\mathrm{M} = (\mathbf{A} - \bar{\mathbf{A}})'(\mathbf{A} - \bar{\mathbf{A}}) / (n - 1).$$

Each element of $\mathbf{A} - \bar{\mathbf{A}}$ is the corresponding element of $\mathbf{A}$ minus the mean over that column. Determining the mean reduces the degrees of freedom by one, hence division by $n - 1$. The analysis the proceeds as described in sections 2.2 and 2.3, using r_m to denote the number of significant factors associated with the distribution about the mean. Returning to the absorbance example, the effect of this manipulation would be to remove the influence of any species absorbing at wavelength j whose concentration is essentially unaltered over the n conditions. Similarities or differences in the estimated ranks can provide powerful constraints on the form of the theoretical model one ultimately wishes to fit to the data.

3. Results

3.1. Previously analyzed data

Several examples from the literature have been tested by this procedure. Comparisons of r_o and r_m (this work) with previous estimates of the rank are given in table 1. To indicate the scope of problems that have been examined in this fashion, the examples are described cursorily. Examples 1 and 2 are absorbance data for methyl red and methyl orange, respectively, at four wavelengths and four pH values, while example 3 comprises 25 such measurements on a mixture of methyl red and methyl orange. Examples 4 and 5 are fluorescence intensities for human serum albumin and sodium naph-

Table 1
Factor analysis of examples from the literature

Example	Array size	Significant factors		Rank
		r_o	r_m	
1	4×4	2	1	2 [b]
2	4×4	2	1	2 [b]
3	5×5	4	2	4 [b]
4	3×3	2	2	2 [c]
5	4×4	1	1	1 [c]
6 [a]	4×10	2–4	2–4	≥4 [d]
7	4×11	2	2	2 [d]
8	5×8	2	2	–
	3×8	2	2	2 [d]
9 [a]	8×8	4	4	3–4 [e], 2 [g]
10	5×6	2	1	2–3 [f], 2 [g]

[a] See text for discussion.
[b] Wallace [1].
[c] Weber [2].
[d] Ainsworth [3].
[e] Wallace and Katz [7].
[f] Ainsworth and Bingham [8].
[g] Magar and Chun [5].

thionate, employing an equal number of excitation wavelengths and emission wavelengths. Example 6 is discussed below. Example 7 is from a kinetic study of the reaction between reduced cytochrome oxidase and oxygen (absorbance data at 11 wavelengths and four times). Example 8 is absorbance data for hemoglobin at eight wavelengths and five levels of saturation with oxygen (a 3 × 8 subset was analyzed by the original author). Example 9 is described below. The observations in example 10 are the initial rates of CO binding to hemoglobin with five different fractions of methemoglobin and six values of the pH.

Three particular cases deserve additional comment. Example 6 is table 1 from Ainsworth [3]. The data are the absorbance at four wavelengths of ten different mixtures of four dyes [phenol red, acridine orange, di-iodo (R) fluorescein, and rhodamine B]. Obviously, the rank of the data matrix must be four, yet the analysis says that only two factors are statistically significant (more specifically, the addition of a third factor does not yield a statistically significant improvement in the residual variance). However, attempting to reproduce the data with only two abstract factors results in a residual standard error of ±0.049, well in excess of what one might reasonably expect, say ±0.003, for the error in the absorbance measurements. Applying the χ^2 test gives a vanishingly small probability that two or even three factors will suffice. No other system gave difficulties of this kind and the source remains obscure, although a probable explanation is given in section 4.

A problem of another kind arises in example 9, from Wallace and Katz [7]. The data here represent absorbance measurements at eight different wavelengths and eight different pH values made on a solution of a pH indicator (methyl red). Previous estimates of the rank of these data have been as high as four, yet factor analysis of the variance about the origin suggested the presence of a fifth factor ($P = 0.039$). Incorporation of the fifth factor reduced the estimated standard error from ±0.0028 to ±0.0014, about half the uncertainty estimated by the original investigators. Analysis of the variance about the mean gave four factors ($P = 0.0011$) with an estimated standard error of ±0.0014. This suggests that the solution being investigated contained a second pH-sensitive chromophore and possibly a small amount of an absorbing species whose concentration did not vary over the pH range investigated.

A slight variant of this problem occurred with example 5, from Weber [5]. The data are relative fluorescence intensities from sodium naphthionate in moderately alkaline solution, using four excitation wavelengths and four emission wavelengths. A rank of one is expected yet factor analysis showed a second factor to be significant ($P = 0.035$). This conclusion is readily rejected as an artifact, perhaps arising from a peculiar error distribution, by noting that the concomitant estimated residual standard error is about one tenth the least significant digit in the data. Hence, the rank is indeed one.

3.2. Dissociation of hemoglobin

Another interesting example is provided by an investigation of the dissociation of tetrameric human hemoglobin to $\alpha\beta$ dimers and (perhaps) to the α and β chains by scanning molecular sieve techniques [6]. See table 2. In this experiment the

Table 2
Gel permeation data for human hemoglobin (weight-average partition coefficients [a])

Gel 1	Gel 2	Gel 3
0.533	0.414	0.329
0.522	0.407	0.315
0.527	0.410	0.316
0.515	0.395	0.305
0.518	0.396	0.308
0.508	0.384	0.298
0.507	0.384	0.297
0.501	0.378	0.294
0.500	0.378	0.293
0.491	0.367	0.284
0.493	0.370	0.286
0.485	0.363	0.278
0.479	0.352	0.274
0.475	0.351	0.271
0.471	0.346	0.268
0.479	0.360	0.272
0.478	0.358	0.272
0.475	0.355	0.270
0.466	0.344	0.262
0.463	0.341	0.260
0.459	0.336	0.257
0.452	0.329	0.245
0.452	0.333	0.248
0.450	0.329	0.245
0.453	0.334	0.252
0.443	0.324	0.241

[a] Estimated uncertainty in all partition coefficients is ±0.0015.

measured quantity is a weight-average partition coefficient, so the matrix **B** of eq. (1) represents the partition coefficients of the m species on the different gels, while **C** comprises the weight fractions at differing total concentrations. Since the weight fractions at a given concentration must sum to one, the rank of **C** equals $m-1$.

Analysis of the data in table 2 resulted in an estimate of one for the rank of the data matrix, suggesting that a single dissociation (to dimer) would suffice to explain the data within experimental error. The relevant covariance matrices are constructed by evaluating

$$v_{ij} = \sum_k a_{ki} a_{kj} / n \qquad \text{(origin)},$$

$$v_{ij} = \sum_k (a_{ki} - \bar{a}_i)(a_{kj} - \bar{a}_j)/(n-1) \qquad \text{(mean)}.$$

For the data of table 2 the unweighted covariance matrices are

$$\mathbf{V}_{\mathrm{O}} = \begin{pmatrix} 0.234 & 0.175 & 0.134 \\ 0.175 & 0.131 & 0.101 \\ 0.134 & 0.101 & 0.077 \end{pmatrix},$$

$$\mathbf{V}_{\mathrm{M}} = 10^{-4} \begin{pmatrix} 7.39 & 7.45 & 6.80 \\ 7.45 & 7.57 & 6.86 \\ 6.80 & 6.86 & 6.29 \end{pmatrix}.$$

Using an estimate of 0.0015 for the uncertainty (standard error) in the data, the corresponding weighted matrices are

$$\mathbf{V}_{\mathrm{O}} = 10^{4} \begin{pmatrix} 10.4 & 7.79 & 5.97 \\ 7.79 & 5.84 & 4.48 \\ 5.97 & 4.48 & 3.44 \end{pmatrix},$$

$$\mathbf{V}_{\mathrm{M}} = \begin{pmatrix} 329 & 331 & 302 \\ 331 & 336 & 305 \\ 302 & 305 & 280 \end{pmatrix}.$$

Scaling the data (weighted analysis) can offer computational advantages besides providing a more immediate grasp of the magnitude of the unexplained variance.

The matrices just given possess the following respective sets of eigenvalues:

$$\begin{array}{lll} (4.42\times 10^{-1} & 7.69\times 10^{-5} & 4.56\times 10^{-6}), \\ (2.12\times 10^{-3} & 4.80\times 10^{-6} & 1.48\times 10^{-6}), \\ (1.97\times 10^{5} & 34.2 & 2.07), \\ (9.42\times 10^{2} & 2.13 & 0.66). \end{array}$$

Applying eq. (4) to the unweighted covariance about the origin

$$F_\alpha(1,2) = 0.44/[(7.7\times 10^{-5} + 4.6\times 10^{-6})/2] = 1.1\times 10^{4},$$

$$F_\alpha(1,1) = 7.7\times 10^{-5}/4.6\times 10^{-6} = 17,$$

when these values are compared with $F_{0.95}$ (1, 2) = 18.51 and $F_{0.95}$ (1, 1) = 161.4, one concludes that the first factor is significant and that the second is not (see table 3). Since the estimate of the residual standard error (±0.006) was much larger than the estimated uncertainty in the data (±0.0015) and there was reason to believe that the dissociation might proceed beyond the dimer, the system was studied more closely.

The first approach was to inflate the rank of the

Table 3
Analysis of hemoglobin data

		$\mathbf{V}_O$	$\mathbf{V}_M$
eigenvalues [a]		1.967×10^5	941.7
		34.22	2.134
		2.075	0.6572
factor 1	SE [a]	4.26	1.18
factor 1	$P(F)$ [b]	9.2×10^{-5}	1.5×10^{-3}
factor 1	$P(\chi)$ [c]	$< 10^{-5}$	0.247
factor 2	SE	1.44	0.81
factor 2	$P(F)$	0.154	0.323
factor 2	$P(\chi)$	0.150	0.417

[a] Scaled by the estimated uncertainty in the data (≈ 0.0015).
[b] Probability that the reduction in residual variance arises from chance.
[c] Probability that the residual variance results from chance.

data matrix by multiplying each partition coefficient by the concentration at which it was determined. This changes **C** from a matrix of weight fractions to a matrix of concentrations and increases its rank from $m - 1$ to m. However only a single significant factor was found, primarily because there is a 100-fold change in concentration but only a 10% change in partition coefficient. The second approach was to augment the original data matrix by including an additional column comprised of the average of the partition coefficients determined at each concentration. This procedure increases the size of the array but leaves the rank and degrees of freedom unaltered. Again there was only one significant factor.

4. Discussion

For many individuals the strongest argument in favor of abstract factor analysis may well be that the necessary programming already exists in standard statistical packages available at most computing centers. However, there are more meaningful reasons for its use. Foremost is the capability of performing a statistically valid test in the absence of prior information about the error distribution. Even under these limiting conditions the test embodied in eq. (4) is capable of estimating the minimum number of species required to describe the data. The application of eq. (5) then provides an estimate of the actual uncertainty in the data. In the event that this estimate is unreasonably small, it means that too many factors have been taken as significant, most probably because the error distribution is nonuniform or nonrandom. Accordingly, the estimate of the rank is reduced.

In the fortunate circumstance that information is available on the magnitude of the uncertainties, their distribution, or both, one can apply a powerful combination of two statistical tests. First, as before, the F-test of eq. (4) assesses the significance of the kth factor by testing the improvement in the variance. Second, one can test the magnitude of the residual variance as a reduced χ^2.

$$\chi^2(p-k) = \sum_{k+1}^{p} \lambda_j/(p-k).$$

One then seeks that value of k which yields a large F_α (small probability that the improvement is due to chance) and simultaneously a value of χ^2_ν which approximates unity ($\approx 50\%$ probability that the residual variance is due to chance). Too large a value for χ^2_ν suggests that additional significant factors remain, whereas too small a value implies that nonsignificant factors (pure error) have been incorporated.

Estimating the number of significant factors by counting the number of eigenvalues greater than one has been suggested in the past as a rough and ready guide to the rank of the matrix. This has a strong dependence on the error estimates, as does the χ^2_ν analysis. In addition, however, this approach supposes that the error distribution is uniform, with structureless error factors. Several of the cases analyzed here had one or more statistically insignificant abstract factors associated with eigenvalues greater than one. In other instances significant factors may be associated with eigenvalues less than one, particularly if the uncertainty in the data has been overestimated. When example 9 was tested by this procedure (5), using the correlation matrix rather than the variance matrix, the estimated rank was only two instead of four. This criterion is not well-grounded statistically and there is little reason to recommend its use.

Since both the data matrix and the error matrix span p-spaces and the solution spans an m-space, separation of the abstract factors into significant and nonsignificant groups extracts that amount of error contained in the $(p-m)$-space. Denoting by $\mathbf{F}_{(m)}$ the matrix obtained by setting f_{ij} equal to zero for all $j>m$, it is then possible to form an estimate of the data matrix with the error factors removed.

$$\hat{\mathbf{A}} = \mathbf{F}_{(m)}\mathbf{F}'_{(m)}\mathbf{A}.$$

(This procedure was used in the numerical calculations to verify that calculations on the eigenvalues did indeed reflect the actual residual variances.) Although the idea has not been tested, in some applications it may prove useful to perform subsequent analyses on the improved estimate $\hat{\mathbf{A}}$ rather than on the raw data $\mathbf{A}$ directly. It then becomes interesting to estimate the error still contained within $\hat{\mathbf{A}}$ (i.e. that error remaining from the m error factors not extracted). By assuming the distribution of error factors to be uniform, this estimate is just the standard error of eq. (5) multiplied by $(m/p)^{1/2}$. Since the distribution of error factors is generally found to be structured, implying that the hidden error factors may well be larger, this estimate must be taken as a lower bound to the remaining error.

Finally, it is well worth repeating a possible pitfall of rank analysis. The estimated rank of $\mathbf{A}$, $r_{\mathbf{A}}$ may be less than m if either $r_{\mathbf{B}}$ or $r_{\mathbf{C}}$ is less than m. This situation will occur whenever one or more of the columns of $\mathbf{B}$ or the rows of $\mathbf{C}$ can be expressed as a linear combination of the others. More formally,

$$\text{if } \exists\,\{\alpha_1,\dots,\alpha_m\} \ni \sum_k \alpha_k c_{ik} = 0 \quad (\text{all } i)$$

then $r_{\mathbf{A}} = r_{\mathbf{C}} < m$, and

$$\text{if } \exists\,\{\beta_1,\dots,\beta_m\} \ni \sum_k \beta_k b_{kj} = 0 \quad (\text{all } j)$$

then $r_{\mathbf{A}} = r_{\mathbf{B}} < m$.

That is, either classical rank analysis or factor analysis will return a minimal estimate of the number of linearly independent species actually present. In some cases the experimenter will have available additional information requiring that more species be considered in subsequent parameter estimation. In other instances it may prove difficult to distinguish between a rank of p and a rank of zero. In the analogy of polynomial fitting, one choice is a line which passes exactly through all points and the other is a point (the mean). These are equally valid interpretations of an array of random numbers. What appears to have happened in example 6 of table 1 is that two linear combinations of the component spectra can account for most of the variance. Perhaps this arises from a limited range of compositions making a third factor in concentration nonsignificant. This example illustrates the importance of overdetermining the system.

It became necessary to consider the question of linear dependence when initial analysis of the data in table 2 resulted in an estimate of one for the rank. Although the partition coefficients for the individual species on the different gels are related to one another in a nonlinear fashion [9], there remained the possibility that within experimental error they were linearly correlated. That is, most of the data come from the region where the relation between log (molecular weight) and partition coefficient is reasonably linear for each gel, although the individual slopes and intercepts would differ. The result of this common linearity would be to make the rank of the matrix of partition coefficients be one no matter how many species were actually present. Such a finding would certainly diminish the utility of gel permeation techniques in studying self-associating systems, so it became important to obtain a more definitive answer. Partition coefficients for six different nonassociating proteins on four different gels were taken from ref. [10]. Analysis of the 6×6 variance–covariance matrix showed the presence of four significant factors. Hence, the permeation properties of the different gels are indeed linearly independent. This is direct experimental confirmation of the previous theoretical deduction [9].

This means that the rank of one is to be assigned to a dissociation and that two species (tetramer and dimer) will be statistically sufficient to describe the data. However, the residual uncertainty is about four times the estimated uncertainty in the data. Several explanations can be offered. First, these two estimates of the uncertainty are not exactly equivalent. The estimated uncertainty in the data is a measure of the reproducibility at fixed concentration, whereas the residual uncertainty in the factor analysis also incorporates errors in concentration. Second, it is entirely possible that some dissociation to monomeric α and β chains does occur, although to an extent that does not provide a statistically significant factor. The kind of additional information that lies outside the scope of factor analysis yet could resolve this question would include a measured weight-average partition coefficient exceeding that of the dimer. Third, there may well be nonideality effects in the gel permeation experiments at higher concentration. If these effects differ among the gels they can appear as additional factors or as elevated estimates of uncertainty. It is certainly possible that all three effects are contributing. It is interesting that the $\mathbf{V}_M$ analysis appears to eliminate this contribution to the variance, suggesting that there might be either a noninteracting species or a small amount of an interacting species occurring at a relatively constant level. Model-dependent calculations are necessary to resolve the source.

Lest it appears that the limitations of factor analysis have been unduly stressed, the advantages are now repeated. The principal power resides in the parsimony of assumptions about the system. One obtains a statistically valid estimate of the minimum number of linearly independent species present without invoking any model. Simultaneously there results an estimate of the residual uncertainty, independent of prior error analysis. This information, limited as it may be, provides powerful constraints on subsequent model-dependent calculations. The ready availability of packaged programs for factor analysis recommends this as a most useful preliminary step in the reduction of experimental data.

Appendix

In doing the calculations reported here I used a collection of reasonably standard subroutines for matrix multiplication and eigenvalue determination, operating in a laboratory minicomputer (16 K words). The individual not interested in preparing his own program will find the necessary features in packaged statistical programs, such as Factor Analysis, program BMDP-P4M developed by the Health Sciences Computing Facility, UCLA. Guidance from the personnel of the computing center should be sought in preparing the data for input and in selecting appropriate options. Specifically, one wants factor analyses of the covariance matrix about the origin and of the covariance matrix about the mean. One section of this program's output provides the eigenvalues of these matrices, which can then be analyzed using the equations given in the text.

Acknowledgement

It is a pleasure to thank Dr. David Smith for a most helpful conversation concerning, among other things, the concept of degrees of freedom. This work was made possible by receipt of a grant GM 23302 from the National Institutes of Health.

References

[1] R.M. Wallace, J. Phys. Chem. 64 (1960) 899.
[2] G. Weber, Nature 190 (1961) 27.
[3] S. Ainsworth, J. Phys. Chem. 65 (1961) 1968.
[4] M.E. Magar, Data analysis in biochemistry and biophysics (Academic Press, New York, 1972).
[5] M.E. Magar and P.W. Chun, Biophys. Chem. 1 (1973) 19.
[6] G.K. Ackers, private communication.
[7] R.M. Wallace and S.M. Katz, J. Phys. Chem. 68 (1964) 3890.
[8] S. Ainsworth and W.S.W. Bingham, Biochim. Biophys. Acta 160 (1968) 10.
[9] G.K. Ackers, J. Biol. Chem. 243 (1968) 2056.
[10] H.S. Warshaw and G.K. Ackers, Anal. Biochem. 42 (1971) 405.