Non-Gramian and Singular Matrices in Maximum Likelihood Factor Analysis

Hei-Ki Dong
The Ball Foundation

In some cases, a correlation matrix may be singular because of the multicollinearity in data, and it may become non-Gramian because of computational inaccuracies. In such cases, popular methods of factor extraction, such as maximum likelihood factor analysis, image factor analysis, and canonical factor analysis, cannot be used because of computational difficulties. This article provides a simple heuristic procedure for converting such a matrix into a proper matrix, so that maximum likelihood factor analysis may be performed.

In recent years, the use of unrestricted (exploratory) maximum likelihood factor analysis (UMLFA) (Jöreskog, 1967) has become very popular due to advances in high-speed computer technology. The method has enjoyed a dramatic increase in its popularity in many disciplines of scientific research. A major reason is because most statistical packages now have incorporated an UMLFA program into their software.

UMLFA requires an $n \times n$ symmetric positive definite (i.e., all eigenvalues greater than zero) matrix $\mathbf{R}$ of intercorrelations among n attributes or variables. In some situations, the condition of positive definiteness is not met, so that UMLFA (and other methods such as image factor analysis, Kaiser, 1963, and canonical factor analysis, Harris, 1963) cannot

be conducted. For example, when $\mathbf{R}$ with negative (or zero) eigenvalue(s) is input to an UMLFA program, the program terminates because of a negative (or zero) argument when attempting to evaluate log $|\mathbf{R}|$, which is undefined whenever $|\mathbf{R}|$ equals zero or negative, where $|\mathbf{R}|$ is the determinant of $\mathbf{R}$. In general, most of these inappropriate cases come from situations in which there is a high degree of multicollinearity among n variables. In such cases, computational inaccuracies (roundoff errors) become paramount. For instance, a correlation matrix may become non-Gramian (i.e., some of the eigenvalues are negative, which is impossible by the definition of $\mathbf{R}$) because of rounding errors.

Also, in UMLFA there is a tendency to produce Heywood (1931) cases (communalities greater than 1.0), and a tendency to produce small estimates $\mathbf{U}^2$ of uniquenesses, resulting in improper solutions. Small uniquenesses become a problem in UMLFA because the sample correlation matrix $\mathbf{R}$ is rescaled by $\mathbf{U}^{-1}$ prior to its eigenvalue-eigenvector decomposition. A very small u results in a large u^{-1}, causing *shifts* in eigenvectors of $\mathbf{R}$, thus drastically influencing the pattern of factor weights (Dong, 1982). (Small uniquenesses imply that $\mathbf{R}$ is close to being singular, i.e., "ill-conditioned.") Incidentally, the problem of shifting eigenvectors is not unique to UMLFA; it applies to any method when the correlation matrix is rescaled by $\mathbf{U}^{-1}$, for example, image factor analysis and canonical factor analysis.

363

In multiple regression analysis, problems are found that parallel those found in UMLFA. For example, the problem of multicollinearity of predictors in multiple regression implies an ill-conditioned correlation matrix **R** among predictor variables. In this case, it is well-known that regression weights are very unstable. This parallels the shifting of factor weights in UMLFA when **R** is non-Gramian, singular, or ill-conditioned. Hoerl (1959, 1962) and his colleague (Hoerl & Kennard, 1970, 1975, 1976) have provided the empirical as well as theoretical justification for adding a small ridge constant k to the diagonal of **R** to stabilize regression weights (ridge regression). This is equivalent to just adding a k to eigenvalues of **R** with no effect on the pattern of associated eigenvectors. In the UMLFA case, this is equivalent to increasing uniquenesses; it is tacitly assumed that there is less communality or more uniqueness among variables than those actually observed. When **R** is rescaled by more stable uniquenesses, results are stabilized.

A simple *heuristic* procedure (called the ridge factor solution) of adding a constant k to the diagonal of **R** is suggested whenever **R** is non-Gramian or singular (i.e., some of the eigenvalues of **R** are zero) because of multicollinearity among variables and computational inaccuracies, forcing **R** to be positive definite and proper for UMLFA. The procedure is not recommended for situations when the non-Gramian or ill-conditioned matrix is due to missing data or to poor study design and execution. Adjusting the improper matrices to make it computationally proper, in such cases, will not necessarily make the solution substantively appropriate.

The procedure is as follows: (1) a constant k is added to the diagonal of **R**, that is, it is tacitly assumed that there is less communality or more uniqueness among variables than those actually observed, (2) **R** is rescaled so that unities are in the diagonal, and (3) the resulting proper **R** is input to an UMLFA program. Various increments of k are tried until stability of factor weights is achieved. The stability is visually inspected by plotting k versus factor weights of each factor, that is, ridge plots.

An Example

The example was taken from Thurstone (1947, pp. 140–147), which has become a classic in factor analytic research and is known in the literature as the "Thurstone's 20-variable box problem." Data consisted of a hypothetical population of 20 boxes, whose basic factors of X, Y, and Z were given in Thurstone (p. 141). From each box's factor scores, Thurstone (p. 142) computed 20 mathematical functions (see Figure 1) and used these function scores to construct the 20×20 hypothetical correlation matrix. The "true" dimensionality of the factor space was known to be three, and the "true" factor pattern was a simple structure.

Because the original **R** was constructed from 20 boxes, which was not greater than the number of mathematical functions (20), and because of the multicollinearity among functions, **R** must be singular. Moreover, due to rounding errors, the original **R** possessed several negative eigenvalues, that is, **R** was non-Gramian, which by the definition of **R**, was impossible.
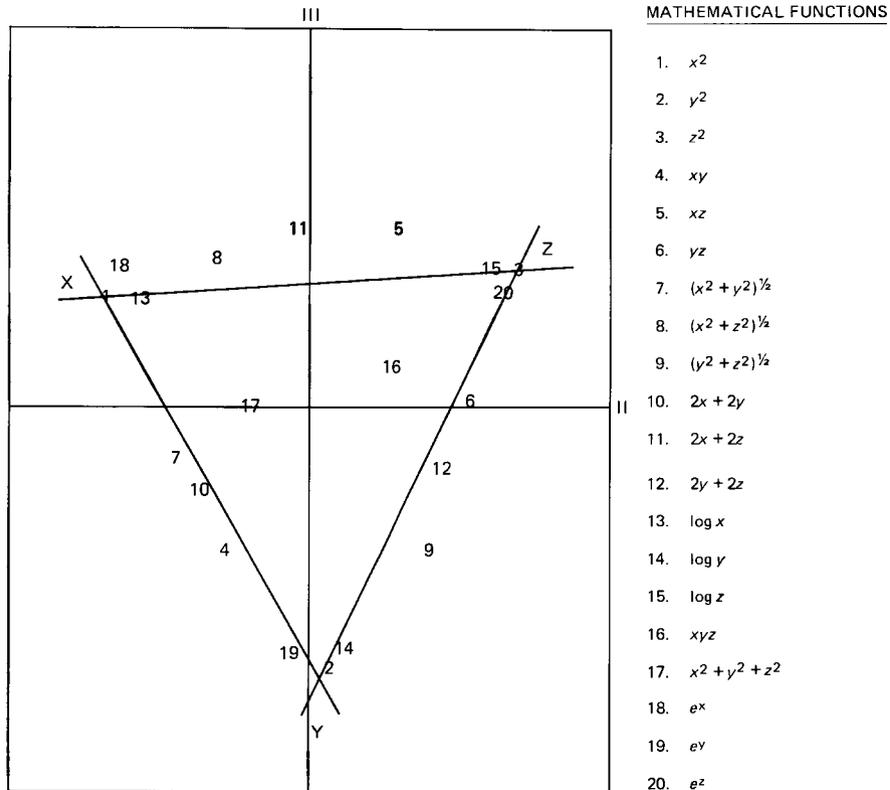
To convert the correlation matrix into a proper (Gramian, positive definite) matrix for UMLFA, a k of .1, .2, and .3 was added to the diagonal of **R**. The resulting matrix was then input to an UMLFA program and analyzed from 1 to 10 factors.

Figure 1 displays the three-factor extended vector projection from UMLFA when $k = .1$. This figure was compared with Thurstone's (1947, p. 147) simple structure figure. The two figures were quite similar, with Factor X defined by functions 1, 13, and 18, Factor Y defined by functions 2, 14, and 19, and Factor Z defined by functions 3, 15, and 20. The rest of the functions lay close to one of the straight lines $\overline{XZ}$, $\overline{XY}$, or $\overline{YZ}$. On which line a particular function lay, depended on what factors were used to compute that function. For example, the function 12 lay on the line $\overline{YZ}$ because it consisted of factors Y and Z, namely, $2y + 2z$. Similar results were obtained for $k = .2$ and .3.

Summary

In summary, when a ridge constant k was added to the diagonal of a non-Gramian (or singular) cor-

Figure 1
Three-Factor Extended Vector Projection from UMLFA
($k = .1$) of Thurstone's 20-Variable Box Problem

MATHEMATICAL FUNCTIONS

1. x^2
2. y^2
3. z^2
4. xy
5. xz
6. yz
7. $(x^2 + y^2)^{1/2}$
8. $(x^2 + z^2)^{1/2}$
9. $(y^2 + z^2)^{1/2}$
10. $2x + 2y$
11. $2x + 2z$
12. $2y + 2z$
13. $\log x$
14. $\log y$
15. $\log z$
16. xyz
17. $x^2 + y^2 + z^2$
18. e^x
19. e^y
20. e^z

relation matrix, converting it to a Gramian (positive definite) matrix, it produced stable and meaningful UMLFA solutions. In general, whenever **R** is singular or non-Gramian, due to multicollinearity among variables and computational inaccuracies, it is recommended that a reasonable ridge constant k of .1, that is, 10% of variance, be added to the diagonal of **R** in order to perform UMLFA.

It should be mentioned that Dong (1982) has successfully applied ridge factor solutions to ill-conditioned matrices to produce stable and meaningful UMLFA solutions. He also demonstrated that when a ridge constant k was added to *proper matrices,* the resulting UMLFA and principal factor analysis solutions did not differ significantly from those with $k = 0.0$, that is, the normal situation.

References

Dong, H. K. (1982). *Ridge constants and maximum likelihood estimation in common factor analysis.* Unpublished manuscript, University of Illinois, Champaign IL.

Harris, C. W. (1963). Canonical factor models for the description of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 138–155). Madison WI: University of Wisconsin Press.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London, 134,* 486–501.

Hoerl, A. E. (1959). Optimum solution to many variable equations. *Chemical Engineering Progress, 55,* 69–79.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress, 58,* 54–59.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12,* 55–67.

Hoerl, A. E., & Kennard, R. W. (1975). Ridge regression: Some simulations. *Communications in Statistics, 4,* 105–123.

Hoerl, A. E., & Kennard, R. W. (1976). Ridge regression: Iterative estimation of the biasing parameter. *Communications in Statistics, 5,* 77–88.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443–482.

Kaiser, J. F. (1963). Image analysis. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 156–166). Madison WI: University of Wisconsin Press.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Acknowledgments

Author's Address

Send requests for reprints or further information to Hei-Ki Dong, The Ball Foundation, 800 Roosevelt Road, Building C, Suite 206, Glen Ellyn IL 60137, U.S.A.