On the Unlikely Case of an Error-Free Principal Component From a Set of Fallible Measures

Educational and Psychological
Measurement

2018, Vol. 78(4) 708–712

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416686147

journals.sagepub.com/home/epm

Tenko Raykov¹, George A. Marcoulides²,
and Tenglong Li¹

Abstract

This note extends the results in the 2016 article by Raykov, Marcoulides, and Li to the case of correlated errors in a set of observed measures subjected to principal component analysis. It is shown that when at least two measures are fallible, the probability is zero for any principal component—and in particular for the first principal component—to be error-free. In conjunction with the findings in Raykov et al., it is concluded that in practice no principal component can be perfectly reliable for a set of observed variables that are not all free of measurement error, whether or not their error terms correlate, and hence no principal component can practically be error-free.

Keywords

error variance, measurement error, observed variable, principal component, principal component analysis, reliability, variance

One of the often used multivariate statistical methods in empirical behavioral and social research is principal component analysis (PCA; e.g., Raykov & Marcoulides, 2008). The main reason for its popularity is the variance optimization property that principal components (PCs) of a given set of manifest measures possess.

¹Michigan State University, East Lansing, MI, USA

²University of California, Santa Barbara, CA, USA

Corresponding Author:

Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443a Erickson Hall, East Lansing, MI 48824, USA.

Email: raykov@msu.edu

Unfortunately, a frequent misinterpretation of this property in applied studies, especially with regard to the first PC, is that this feature implied somehow the possibility of a PC—for instance the first PC—having perfect reliability. This misinterpretation seems to be also related to arguments aiming at supporting claims that the first PC extracts “all reliable information” from an analyzed set of variables, in particular in settings of missing data analysis (cf. Howard, Rhemtulla, & Little, 2012).

Recently, Raykov, Marcoulides, and Li (2017) pointed out that if (a) at least one measure is fallible in a set being analyzed via PCA and (b) the pertinent measurement errors are uncorrelated (when two or more observed variables are not error-free), then no PC can have lower error variance than that in the fallible measure with minimal error variance. An important assumption of their developments was that of no correlated error sources (measurement errors) in the analyzed variables, which seems to be advanced explicitly or implicitly similarly in many empirical studies. It is therefore of interest to find out whether the resulting imperfect reliability feature of any extracted PC—and in particular the first PC—remains valid also when at least two error terms are correlated across fallible measures. The present note discusses this theoretically and empirically relevant setting and shows that up to an unlikely event the imperfection feature of any extracted PC persists then as well.

Background, Notation, and Assumptions

We assume that a set of observed measures is given, whereby at least two of them contain measurement error (with positive variance), as it will often be the case in the majority of contemporary studies in the educational and behavioral disciplines (see Raykov et al., 2016). We denote these measures by $X_1, \dots, X_k$ ($k > 1$) and their collection by Ξ , that is, $\Xi = \{X_1, \dots, X_k\}$. Using the classical test theory decomposition, which effectively always exists in empirical educational, behavioral, and social research (e.g., Raykov & Marcoulides, 2011), $X_j = T_j + e_j$ holds, where T_j and e_j are the true and error score of the j th measure, respectively ($j = 1, \dots, k$). In addition, we assume in the remainder that at least two of these k error scores, e_1 through e_k , are correlated. Last but not least, we assume that any observed variable used below, including all PCs, has positive variance—a condition that can be hardly considered restrictive in the overwhelming majority of contemporary behavioral and social studies.

In PCA, as is well known, particular linear combinations $Y_1, \dots, Y_k$ of the original measures in Ξ are found, which possess the property that the first one, that is, Y_1 , maximizes observed variance, and in addition each of the following PCs $Y_2, \dots, Y_k$ maximize the “remaining” variance in the set of $X_1, \dots, X_k$ under the condition that any subsequent PC is uncorrelated with all preceding PCs extracted before it (e.g., Timm, 2002). As usual, we assume in the rest of this article that the covariance matrix Σ of $X_1, \dots, X_k$ is positive definite and subjected to PCA. (Analyzing instead the correlation matrix does not alter the following developments, findings, and interpretations.)

We stress that the remaining discussion evolves exclusively at the population level, like that in Raykov et al. (2016).

As elaborated in detail in the literature, the weights in the linear combinations of $X_1, \dots, X_k$ that render each of the PCs are the elements of the normalized eigenvectors pertaining to the successive eigenvalues of Σ (e.g., Johnson & Wichern, 2002),

$$\Sigma = P\Lambda P', \quad (1)$$

with the orthogonal matrix P containing these eigenvectors as columns and Λ being the diagonal matrix with these eigenvalues along its main diagonal (priming denotes transposition in this note).

How Probable Is It for a Principal Component to Be Error-Free?

If interested specifically in the j th PC, Y_j , denoting by p_{sj} its pertinent weights for the observed variables, the following equation holds:

$$Y_j = p_{1j}X_1 + p_{2j}X_2 + \dots + p_{kj}X_k \quad (2)$$

($j = 1, \dots, k$). Using the classical test theory decomposition for the manifest measures (note that if the s th of them, X_s , is error-free, then $e_s = 0$ is valid for all cases; $1 \leq s \leq k$), evidently the corresponding decomposition of Y_j is as follows (e.g., Raykov et al., 2016):

$$\begin{aligned} Y_j &= p_{1j}X_1 + p_{2j}X_2 + \dots + p_{kj}X_k \\ &= (p_{1j}T_1 + p_{2j}T_2 + \dots + p_{kj}T_k) + (p_{1j}e_1 + p_{2j}e_2 + \dots + p_{kj}e_k). \end{aligned} \quad (3)$$

From Equation (3) it follows, therefore, that the j th PC will be error-free if and only if the variance of its error score is 0, that is, if and only if

$$\text{Var}(p_{1j}e_1 + p_{2j}e_2 + \dots + p_{kj}e_k) = 0, \quad (4)$$

where $\text{Var}(\cdot)$ denotes variance ($j = 1, \dots, k$). However, in the presently considered setting Equation (4) will hold if and only if

$$p_{1j}^2\text{Var}(e_1) + p_{2j}^2\text{Var}(e_2) + \dots + p_{kj}^2\text{Var}(e_k) + 2 \sum_{q < r} p_{qj}p_{rj}\text{Cov}(e_q, e_r) = 0, \quad (5)$$

where $\text{Cov}(\cdot, \cdot)$ denotes covariance and the sum on the left-hand side (LHS) of (5) runs over all pairs of error terms that are correlated (i.e., over all error term pairs with nonzero covariance that are included only once in it).

From Equation (5) it is seen that a PC will not contain any measurement error if and only if the nonzero error covariances are such that their linear combination with weights being the products of the corresponding PC weights equals (half of) the

negative of the linear combination of the error variances with weights being the squares of the pertinent PC weights.

Given that the j th PC is fixed in the discussion in this section, the variances and covariances in the LHS of Equation (5) span a proper subspace S of R^{k+t} , where S could be viewed as the parameter space for the presently considered setting concerned only with the error variance in Y_j , t is the number of (nonduplicated) nonzero error covariances in (5), and R^{k+t} denotes the set of all real vectors of dimension $k+t$. We note that the space S has also dimensionality $k+t$. The reason is the fact that there is no relationship between the variances and covariances in the LHS of (5) (apart from that resulting from the error variance of Y_j being nonnegative), which could make the dimensionality of S lower than $k+t$. However, the set of all $(k+t)$ -dimensional vectors consisting of possible values for the variances and covariances in the LHS of (5), which satisfy the condition (equality) stated in Equation (5), is a proper subspace S^* of S . This is because the LHS of (5) is a linear combination of all the variances and covariances in it that generated the space S to begin with. For this reason, the dimensionality of S^* is lower than $k+t$, that is, $\dim(S^*) < k+t$, where $\dim(\cdot)$ denotes dimensionality. Therefore, S^* has mass (probability) of 0 in the space S of interest here (e.g., Apostol, 2013).¹

Hence, from the preceding discussion it follows that the probability of the condition in Equation (5) holding for a particular PC, say the j th, is 0 ($j = 1, \dots, k$). That is, in practice it is (in effect) impossible that Equation (5) will hold in a given empirical study (even if conducted at the population level). Therefore, it is practically impossible for a PC, which is extracted from a set of observed measures where at least two are fallible, to have perfect reliability regardless whether the error terms of these fallible measures correlate (for the case of uncorrelated errors, see Raykov et al., 2016; see also Note 1).

Conclusion

The impressive popularity of PCA in the empirical sciences may have contributed to the circulation of some myths about PCs among certain quarters of applied researchers. One of them is that the first (or subsequent) PC extracts “all reliable variance” in the analyzed set of observed variables that may—and usually will—contain imperfect measures in educational and psychological studies, with the potential implication that at least the first PC may be perfectly reliable. The present note extended the earlier discussion in Raykov et al. (2016) dealing with the uncorrelated error case, to the setting where at least two errors in fallible observed variables possess nonzero covariance. The note demonstrated that while it is possible for the first (or later) PC to have perfect reliability in case of correlated errors—when a set of manifest measures is subjected to PCA and at least two of them are fallible—the probability is 0 of this occurring in a given empirical study (even if conducted at the population level; see also Note 1).

Therefore, it is practically impossible for a PC to be error-free if extracted from a set of observed measures with at least two of them containing measurement errors that correlate. Together with the findings in Raykov et al. (2016) this note leads thus to the conclusion that in practice no PC can be free of measurement error for a set of observed variables that are not all perfectly reliable, irrespective of whether the error terms of the individual measures correlate, and hence no PC can practically be free of measurement error then. With this conclusion in mind, the present note's message is therefore that while PCs have highly desirable, observed variance maximization features, in practical terms this does not imply that a PC—even the first—will or could be error-free if not all measures in the PC analyzed set are perfectly reliable.

Acknowledgment

Thanks are due to S. Penev for a valuable discussion on principal components.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. This reasoning, as well as that underlying all developments in Raykov et al. (2016), is based on the fact that an observed variable weight in a principal component under consideration can be 0 only with probability 0 (cf. Apostol, 2013). That is, all no error-free principal component and related statements in this note as well as in Raykov et al. (2016) are correct with probability 1, that is, in practical terms always.

References

- Apostol, T. (2013). *Calculus*. New York, NY: Springer.
- Howard, W. J., Rhemtulla, M., & Little, T. D. (2012). Using principle components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, 50, 285-299.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.
- Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer.
- Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. New York, NY: Taylor & Francis.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- Raykov, T., Marcoulides, G. A., & Li, T. (2017). On the fallibility of principal components in research. *Educational and Psychological Measurement*, 77, 165-178.