Multivariate Behavioral Research, 27 (2), 219-224

Extension of Guttman's Result From *g* to PC1

Peter H. Schönemann
Purdue University & National Taiwan University

In his trenchant critique of Jensen's (1985) "house of cards" (p. 202), Guttman noted that "... more algebraic thinking ... might have prevented him from dismissing out of hand any suggestion, made in some peer comments, that the second hypothesis [of positive Spearman, 1927, correlations] might be but an algebraic consequence of the first [that *g* exists]" (p. 198).

Given the cosmic scope of Jensen's (1985) visions — "implications ... for employment, productivity and the nation's welfare" (p. 206) — one might also have wished for less latitude in his varied verbal definitions of "Spearman's hypothesis" which, as it soon turned out, leave room for two different technical interpretations: (a) a "Level I version" — the mean difference vector **d** correlates positively with the regression weights of the first principle component (PC1) of the *pooled* correlation matrix; and, (b) a "Level II version" — the mean difference vector **d** correlates positively with the regression weights of the PC1s of *both within* correlation matrices" (Schönemann, 1986, p. 1).

Although they have different implications, Jensen (1985) used both interpretations interchangeably: In Jensen (1980), he appealed to the weaker Level I version when he pointed to "Probably the most compelling assemblage of evidence for the Spearman hypothesis, from the standpoint of factor analysis, ... the massive data of the General Aptitude test of the US Employment Service" (p. 549). For these data, he reported a correlation of .71 between the mean white/black difference vector and the *g* loadings of the *pooled* white/black sample (Jensen, 1985, p. 216). These correlations will be called *Spearman correlations* from now on. In Jensen (1985), on the other hand, he appealed to both versions simultaneously when he offered, in addition to the GATB data (Level I), other data sets for which the mean difference vectors correlated positively with both *within sample* principle components (Level II).

Not surprisingly, perhaps, the positive Spearman correlations remain artifacts under both interpretations.

This article was completed while the author held a visiting research professorship at the National Taiwan University, Republic of China. Support by the National Science Council of the Republic of China is greatly acknowledged.

In Schönemann (1985), I showed that they are artifacts under the Level I interpretation: "if [the mean difference vector] **d** is chosen large enough, it will approximate the largest eigenvector of the pooled covariance matrix **C**" (p. 241). This will be true on Level I whether all within covariances are positive or not.

When Shockley (personal communication, November 6, 1986) correctly questioned the relevance of my Level I argument for the stronger Level II interpretation, I extended it to Level II in an article which I submitted on November 26, 1986, to *The Behavioral and Brain Sciences*:

> To extend this reasoning to the within sample case, one needs a positive manifold (which implies a dominant first eigenvector with equal signs) and the assumption that the pooled distribution is *approximately* multivariate normal, ... Then any roughly equal split into a HI and LO group produces two attenuated within covariance matrices whose principle components will be *approximately parallel* to the principal component of the pooled sample. Hence the first eigenvector will *correlate highly* with the mean difference vector in all three samples (Schönemann, 1986, p. 2, emphasis added).

The editor rejected this manuscript on the advice of an Associate Editor who wrote: "Schonemann's [comments] are mostly hollow and I do not believe warrant publication ... he can exhibit a special case (mathematically) where a positive correlation exists. *But we already know from Jensen's data that such a positive correlation can exist, and more interestingly, does evidently exist with real data*" (Rubin, 1986). I have added the emphasis to the curious "real data" logic: Why would anyone care if the Spearman correlations had never arisen with real data?

In the target article, Guttman proved the sharper result of *perfect* collinearity of **d** with all three *g*s under the stronger assumptions (a) that Spearman's (1927) factor model holds in both subpopulations and (b) also in the pooled total population.

Although Jensen (1985) invokes these assumptions routinely when he extracts his PC1s and then talks about them as if they were *g*, Guttman, of course, knew that in practice Spearman's factor model virtually never fits the data for only one factor, *g*: "Any reader of these lines can himself easily disprove *g* by looking at almost any mental test correlation matrix at his disposal and checking for proportionality" (p. 182).

It should therefore be of interest that a similar result can be derived for principal components (rather than *g*) without any need to invoke the unrealistic factor model: All one needs is (a) that the total (pooled) distribution is

multivariate normal, and (b) that all subtest correlations remain positive in both subpopulations (defined by a bisecting plane which contains the centroid).

In this case, positive Spearman correlations emerge as artifacts on Level II because, as will now be shown, the mean difference vector for the High and Low subpopulations will be *perfectly collinear* with the PC1s of all three covariance matrices. The argument needed to show this is essentially just an algebraic reformulation of the geometric argument which had failed to convince Professor Rubin in 1986.

Main Result

Theorem

If the range Re^p of a p-variate normal random vector $\mathbf{y} \sim N_p(0,\Sigma)$ is partitioned into a High set (H) and a Low set (L) by the plane $\Sigma_k \mathbf{y}_k = 0$, and both within covariance matrices Σ_H, Σ_L remain positive, then (a) the mean difference vector $\mathbf{d} = E(\mathbf{y}|H) - E(\mathbf{y}|L)$ is collinear with the largest principal components of Σ, and (b) $\mathbf{d}$ is also collinear with the PC1s of Σ_L, Σ_H.

Proof

Let $\mathbf{y} \sim N_p(0,\Sigma)$, where Σ is positive and has eigen-decomposition $\Sigma = SD^2S'$, with the eigen-values c_k^2 ordered by magnitude. Then rotation with S, $v = S'\mathbf{y}$, brings the ellipsoidal density into principal axes position with

$$v_k \sim n(0, c_k^2), \quad k = 1, p.$$

The plane $P: = \Sigma_k \mathbf{y}_k = 0$ defines $H = (\mathbf{y}|\Sigma_k \mathbf{y}_k > 0)$, and L as its complement. Let v_1 be its first principal component (PC1) with variance c_1^2. Then $v_1/c_1 \sim n(0,1)$ implies the conditional means

$$E(v_1|H) = [n(0) - n(\infty)]/[N(\infty) - N(0)] = 2n(0) = 2(2\pi)^{-1/2} = -E(v_1|L),$$

while

$$E(v_k|H) = E(v_k|L) = 0 \quad \text{for } k > 1,$$

since $N_p(0,\Sigma)$ is symmetric with respect to v_1. Therefore, in the principal axes frame the mean vector is

$$[2c_1(2\pi)^{-1/2}, 0, ..., 0) \quad \text{in H.}$$

Since the mean vector in L is its reflection about p, the mean difference vector is

$$\mathbf{d}_v' = [4c_1(2\pi)^{-1/2}, 0, ..., 0).$$

On rotating it back into the original variable frame, the mean difference vector becomes

$$\mathbf{d} = S\mathbf{d}_v = 4c_1(2\pi)^{-1/2}\mathbf{s}_1,$$

where s_1 is the eigenvector associated with the largest root c_1^2 of Σ. Hence $\mathbf{d}$ is collinear with the first principal component of Σ, which proves the first part, (a).

To prove the second part, (b), note first that both within covariance matrices are equal,

$$\Sigma_H = \Sigma_L = \Sigma_W$$

since $N_p(0,\Sigma)$ is symmetric about the plane P. Now consider the conventional sums of products breakdown,

$$\mathbf{T} = \mathbf{W} + \mathbf{B},$$

of the total sums of products matrix $\mathbf{T}$ into a sum of the between sums of products matrix $\mathbf{B}$ and the within sums of products matrix $\mathbf{W}$. In the present case,

$$\mathbf{T} = (N\text{-}1)\mathbf{C},\ \mathbf{W} = (N/2\text{-}1)\mathbf{C}_H + (N/2\text{-}1)\mathbf{C}_L,\ \mathbf{B} = N\,\hat{\mathbf{d}}\hat{\mathbf{d}}'/4,$$

where $\mathbf{C}$ denotes the total covariance matrix estimate, $\mathbf{C}_H$, $\mathbf{C}_L$ the two within covariance matrix estimates, $\hat{\mathbf{d}}$ the estimated mean difference vector, and N the total sample size. Hence $\mathbf{T}$ has N-1 df., $\mathbf{W}$ has N-2 df., and $\mathbf{B}$ has 1 df. Since

$$E(\hat{\mathbf{d}}\hat{\mathbf{d}}') = 4\Sigma_W/N + \mathbf{d}\mathbf{d}' \Rightarrow E(\mathbf{B}) = \Sigma_W + N\mathbf{d}\mathbf{d}'/4$$

one finds, on taking expected values,

$$E(\mathbf{T}) = (N\text{-}1)\Sigma = (N\text{-}2)\Sigma_W + \Sigma_W + N\mathbf{d}\mathbf{d}'/4 \Leftrightarrow \Sigma_W = \Sigma - N\mathbf{d}\mathbf{d}'/4(N\text{-}1).$$

Since $\mathbf{d}$ is an eigenvector of Σ and $\mathbf{d}\mathbf{d}'$, it is also an eigenvector of Σ_W. In particular, if Σ_W is positive, as assumed, then $\mathbf{d}$ is the unique eigenvector

associated with the largest latent roots of Σ, Σ_H, and Σ_L, (Perron's Theorem), which proves (b).

Q.E.D.

Discussion

It thus emerges that the *cosines* between the mean difference vector and the PC1s of the total and the two within-sample covariance matrices are not just positive in general, but except for sampling error, are unity. This fact is obscured when the analyses are based on correlation matrices, or when collinearity is measured in terms of correlations instead of cosines.

More importantly, the present argument does not require the a priori unrealistic assumption that all three covariance matrices satisfy the factor model for one common factor (g).

On the other hand, perfect collinearity is tied to the equal split by the partitioning plane p. If it is translated away from the centroid, the Σ_H will no longer equal Σ_L and the likelihood increases that Σ_H may no longer be positive so that Perron's theorem no longer applies. However, the above geometric argument implies at once that in this case the Spearman correlation will be largest for the larger sample, because (a) the covariance matrix estimate will be less accurate for the smaller sample, and (b) the joint distribution of the smaller sample will have smaller eccentricity. This provides an empirical test. On checking, these predictions were borne out in simulations and also for those of Jensen's (1985) data which involved uneven splits (Schönemann, 1986, 1988, 1989).

Numerous commentators on Jensen's (1985) article greeted his Spearman correlations enthusiastically as confirmation of the black inferiority myth. Given past events, one might have hoped for greater sensitivity to the pernicious implications of such shallow reasoning. Most psychologists can at least plead lack of formal training as a plausible excuse. It is more difficult to fathom what moves noted statisticians to implicitly endorse Jensen's absurd claims by blocking valid refutations.

References

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1985). The nature of black-white differences on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8*, 193-263.

Rubin, D. B. (1986). *Review of Schönemann (1986)*. Manuscript submitted for publication.

Schönemann, P. H. (1985). On artificial intelligence. *Behavioral and Brain Sciences, 8*, 241-242.

Schönemann, P. H. (1986). *Level I and Level II artificial intelligence.* Manuscript submitted for publication.

Schönemann, P. H. (1988). Spearman's hypothesis and the General Toy Factor (Technical Report Series No. 88-2). West Lafayette, IN: Purdue University Mathematical Psychology Program.

Schönemann, P. H. (1989). Some new results on the Spearman hypothesis artifact. *Bulletin of the Psychonomic Society, 27,* 462-464.

Spearman, C. (1927). *The abilities of man.* New York: McMillan.