A critique of statistical methods used to adjust for Individual Quality in order to un-mask underlying trade-offs

Jeffrey A. Walker *

August 17, 2015

1 Introduction

In a series of related papers, Van Damme et al. (2002) and Wilson et al. (2014) observed only positive correlations between an array of athletic performance traits in humans and suggested that performance trade-offs are masked by individual quality. That is, variation in "health, fitness, nutrition, and perhaps other developmental or genetic effects" (Wilson et al., 2014) overwhelm the contribution to performance trade-offs that arise from biomechanical and physiological constraints. Wilson et al. (2014) argue that variation in quality among individuals masks underlying functional trade-offs in a similar way that interindividual variation in the acquisition of resources can mask trade-offs between life history traits if there is large interindividual variation in resource acquisition (Van Noordwijk and De Jong, 1986). Both Van Damme et al. (2002) and Wilson et al. (2014) argue that the correct pattern of performance trade-offs can be recovered by using multivariate procedures to adjust for (or "control") individual quality. Following these mathematical adjustments, the expected negative correlations emerged. Here, I show that inferences from both studies are compromised by severe methodological errors.

Much of the background for this work is in Walker and Caddigan (2015), which develops a model of how functional trade-offs are masked by individual quality. This model uses the graphical algebra of path models, which were specifically developed by Sewell Wright (1918; 1932; 1934) to model the underlying factors ("causes") generating correlations among measured traits. In addition, Walker and Caddigan (2015) provide a methodologically correct method to statistically adjust for something like Individual Quality. Here, I make extensive use of these path models to show the mathematical or inferential errors in Van Damme et al. (2002) and Wilson et al. (2014). The scripts for generating simulated data using all of the path models introduced below are available on GitHub at `https://github.com/middleprofessor/NCAA_decathlon`.

*Address: Department of Biological Sciences, University of Southern Maine, 96 Falmouth St., Portland, ME, 04103, U.S.A. E-mail: walker@maine.edu

2 Statistical problems with previous attempts to adjust correlations to control Individual Quality

Four different types of statistical analyses were used to estimate quality-free correlations. All resulted either in spurious correlations or correlations that are not "quality-free" or free of "inter-individual effects." The errors result from misconceptions of PCA and multivariate statistics.

2.1 Method 1: Data culling

Van Damme et al. (2002) analyzed performance in the ten events of the decathlon from 600 elite athletes and found only positive values in the matrix of pairwise correlations between the ten events. Van Damme et al. (2002) noted that the expected negative correlation between the 100 m and the 1500 m only emerged after limiting the analysis to the top performers (those with total scores over 8,000 points). Van Damme et al. (2002) suggested that this comparison of athletes of similar over-all ability controls for athletic quality. Garland Jr (1994) showed the error in this method. By limiting the analysis to only the top performers, Van Damme et al. (2002) effectively culled variation along the major axis of the bivariate scatter since decathlon scores and the scores on the major axis will be highly correlated (the culling method is similar in spirit to the PC regression residual method below). The necessary consequence of this culling is to shift the correlation of the remaining data in a negative direction. I illustrate this with a simple example of two traits (call them performance traits P_1 and P_2) that are positively correlated because they share a common factor Q. Specifically, 5000 points were generated using the equations

$$\begin{aligned} Q &= \mathcal{N}(0,1) \\ P_1 &= \sqrt{0.5}Q + \sqrt{0.5}\mathcal{N}(0,1) \\ P_2 &= \sqrt{0.5}Q + \sqrt{0.5}\mathcal{N}(0,1) \end{aligned} \tag{1}$$

where $\mathcal{N}(0,1)$ is a random variable drawn from a normal distribution with zero mean and unit standard deviation. The expected correlation is .5 while the actual correlation is .53 (Fig. 1A). In Fig. 1B, the 2500 points (50% of the data) with the smallest scores on the major axis have been removed and the correlation has shifted to .08. In Fig. 1C, the the 4750 points (95% of the data) with the smallest scores on the major axis have been removed and the correlation has shifted to -.42. As the culling moves toward 100%, the correlation will move to -1 (and will be -1 at the limit). For the decathlon data, a negative correlation between culled 100 m and 1500 m performances must emerge if the data are culled enough, which they were once Van Damme et al. (2002) limited the data to athletes scoring above 8000 total points. The culled correlation contains a large negative component that is an artifact of computing a correlation from data culled in this manor. This does not mean that there is not a trade-off between the 100 m and 1500 m events that is being masked by something like Individual Quality. I am simply saying that the result (a negative correlation between the remaining data after severe data culling) is guaranteed.

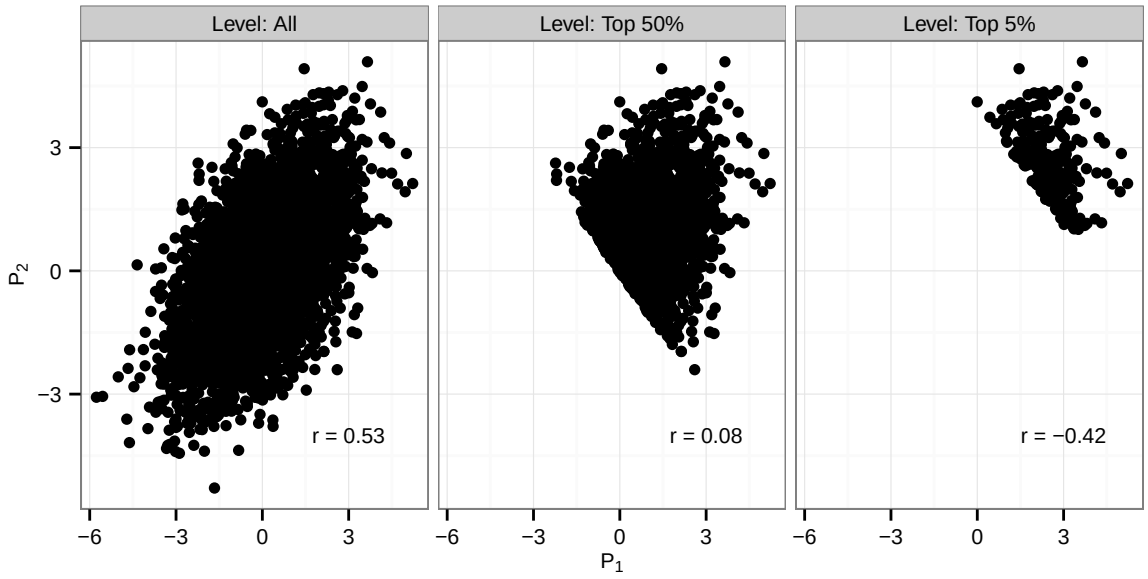
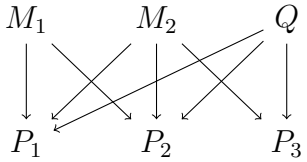


Figure 1: **Culling individuals along the major axis of a bivariate distribution has the effect of moving the subsample correlation toward the opposite of the whole-sample correlation.** The data for 5000 individuals were generated to have an expected correlation of $r = .5$. The sample correlation is $r = .53$ (left panel). Limiting the analysis to the individuals with the top 50% overall score (PC1) shifts the correlation to $r = .08$ (middle panel). Further limiting the analysis to the individuals with the top 5% overall score (PC1) shifts the correlation to $r = -.42$ (right panel).

2.2 Method 2: Partial correlation conditional on all other performances

Van Damme et al. (2002) used the data culling method only for the 100/1500 meter comparison. For each of the 45 correlations between all pairs of the ten decathlon events, Van Damme et al. (2002) computed the partial correlation conditional on all other events. A partial correlation between variables P_i and P_j conditional on the set of variables P_k (which can be one or more variables) measures the part of the covariance between P_i and P_j that is not shared with any of the P_k. It is the correlation of the residuals of P_i and P_j after removing the linear dependence of each on P_k. While we might want to condition on Q (see below), we do not want to condition on all other performance traits because this effectively removes part of the covariation that we want to measure (that is, part of the covariation reflecting the functional architecture of the form-function map). To show this, I use the path diagram of two M-P traits that cause functional correlations among three performance traits. In addition, a quality trait (such as hours training) contaminates the correlations.



(2)

P_1 and P_2 share covariance due to M_1 and M_2 but the part due to M_2 is also shared with P_3. If we want to estimate the correlation between P_1 and P_2 as if there were no contribution from Q, then we want to estimate the net functional correlation $\beta_1 1 \beta_1 2 + \beta_2 1 \beta_2 2$. But the partial correlation of P_1 and P_2 conditional on P_3 partials out the covariance due to M_2, in addition to that due to Q. It over-partializes. Mitteroecker and Bookstein (2009) provide an excellent introduction to this mistake in the context of morphometrics. A partial correlation between P_1 and P_2 conditional on Q (but not the 3rd P) would seem to be reasonable method for estimating the net functional correlation but the result is not scaled properly; it is the partial covariance that estimates the net functional correlation, although the estimate is biased (see below). In summary, a partial correlation between two performance variables conditional on all remaining performance variables is not biased because it is not an estimate of the net functional correlation, but that is the problem, it is not an estimate of the net functional correlation. A partial covariance between two performance variables conditioned on a quality variable is an estimate of the net functional correlation, but it is biased.

2.3 Method 3: Regression residual correlations

To specifically test the Individual Quality hypothesis, Wilson et al. (2014) measured five performance traits related to power, acceleration, and endurance among semi-professional, male soccer players and, like Van Damme et al. (2002), found positive correlations among all traits. Wilson et al. (2014) used the first principal component scores of the correlation matrix as an estimate of Q and "controlled" for Q by computing the correlations of the residuals of the performance traits regressed on Q. The resulting correlation matrix is the set of partial correlations conditional on Q mentioned above except that Q is an unmeasured, latent variable and not a measured variable like "training hours".

There are two statistical problems with using these regression residual correlations as estimates of the net functional correlations. The first is, perhaps, not-intuitive but easily solved. This is, the estimate of the intrinsic performance correlation (as a residual from the total correlation) is properly scaled by computing the *covariance* of the residuals and not the correlation. This covariance estimates the correlation because the residuals of the variance-standardized performances on PC1 of the correlation matrix are already scaled correctly. The variance of these residuals, however, is less than one, so the covariance and correlation matrices among the residuals are not equivalent. But it is the covariance matrix that contains the properly scaled estimates of the quality-free correlation.

Second, even if scaled properly, the estimates have a large, negative bias (Walker and Caddigan, 2015). This bias is not commonly recognized in biology but has been known since at least Karl Pearson (Aitchison, 2003), who invented principal components analysis, and Sewell Wright (Wright, 1932), who invented path models. The geometry of this bias is easily visualized with a bivariate dataset (the raw correlation matrix contains only two variables) and the recognition that the residual values from the regression on the major axis scores is equivalent to the values obtained by sliding the points parallel to the major axis until that they all lie on a single axis normal to the major axis (that is, they all have the same major axis score, i.e. PC1 has been "controlled"). This sliding is similar to the data-culling above (Fig. 1) except that all data are retained and the variance on the major

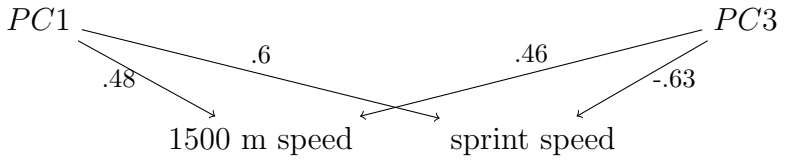
axis is compressed to zero. Given that there are only two axes of variability in bivariate data, controlling variation on one axis necessarily means there is only a single axis of variation remaining. So if the original correlation between the two performance traits is positive, the correlation of their major-axis adjusted residuals must be -1. More generally, if there is only a single factor Q affecting each of p performance variables (that is, there are no M-P traits M as common causes), then the expected correlation among the residuals of each each performance trait regressed on the major axis will be $-\frac{1}{p-1}$. For correlations in which there are multiple M-P traits in addition to Q, the adjusted correlations will be biased in both positive and negative directions (Walker and Caddigan, 2015). These correlations are exceedingly poor estimates of the intrinsic performance correlations (Walker and Caddigan, 2015) and, as a consequence, many of the signs will simply be spurious rather than reflecting the true functional architecture of the form-function map.

2.4 Method 4: Pattern of principal component loadings

The final method used to infer underlying trade-offs was the pattern of loadings on different principal components. This method was only a small part of (Wilson et al., 2014) but two of the three reviewers of the original submission of this manuscript advocated this method, which suggests widespread misunderstanding of PCA. In (Wilson et al., 2014), they note that the pattern of loadings on PC2 is "indicative of a negative correlation between acceleration−power with maximum sprinting speeds". One reviewer of our manuscript stated "when one looks at the third dimension of the PCA, which explains 14% of the variance in the data, it describes the negative correlation between performance in the 1500 m event and the 40m sprint." And another reviewer wrote "If one finds negative relations of performance traits on one or more of the PCs, and if that PC can be considered statistically significant, then you have demonstrated a real negative relation beyond any effects of overall variation in size or quality."

First, this misconception leads to an absurd conclusion if applied objectively. In Table 3 of (Wilson et al., 2014), the loadings for 1500 m and 40 m (sprint) speeds are large and of the same sign on PC2 (-.73, -.47) but large and of opposite sign on PC3 (.46, -.63). The functional correlation due to PC2 is .34 while that due to PC3 is -.29. Using this misconceived interpretation, we come to the absurd conclusion that after holding PC1 ("Individual Quality") constant we have both a facilitation and a trade-off between 1500 m and 40 m speeds!

I try to unpack the basis of this misconception with the following path diagram of part of the PCA decomposition, where the path coefficients are the loadings from Table 3 of (Wilson et al., 2014).

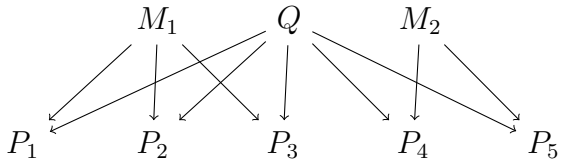


(3)

This path diagram indeed shows a trade-off in the mapping of PC3 to 1500 m and sprint speeds. But it is a misconception to believe that PC3 represents some feature of the

underlying morphology or physiology, that, if tweaked, will cause 1500 m speed and sprint speed to change in opposite directions. It doesn't. We might use the large loadings on PC3 to give it an interpretation. But PC3 neither represents some list of M-P traits nor is it a "real" thing in the sense that it exists independent of this particular decomposition; to believe so is the *reification fallacy*. Instead, PC3 is a weighted sum of the five performance variables; it's a little bit of each one. Nothing about morphology or physiology went into the construction of PC3; we simply cannot learn anything about the functional architecture of the form-function mapping from a table of PCA loadings.

The path model in equation 3 is not a model of causal relationships between M-P traits and performance but a descriptive model of one particular mathematical decomposition of a correlation matrix. Importantly, the patterns of loadings in a PCA are highly constrained by the two fundamental properties of PCA: that each component is determined by the direction of the greatest variation in the space orthogonal to all previous components, and, therefore, that all components are mutually orthogonal. The orthogonality constraint guarantees that components after the first have loadings of opposite sign. I show this with a simulated dataset (Supplement 1) in which five performance traits are generated with the following causal structure



(4)

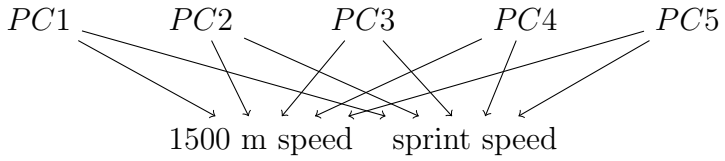
Here, the functional mapping of the M-P traits to performance has a modular organization, that is, each M-P trait causally affects a distinct set of performance variables. The path coefficients from Q to all performance variables are set to .6. The path coefficients from M_1 to P_1, P_2, and P_3 are all set to .4. The path coefficients from M_2 to P_4 and P_5 are set to .3. There are functional facilitations within modules but no functional-trade-offs either within or among modules. The eigenvalues and loadings are given in Table 1 for the first three components. The loadings of opposite signs on PC2 might suggest functional trade-offs between performance variables that occur in different modules. The loadings of opposite signs on PC3 might suggest functional trade-offs between the two performance variables that occur within module two. But of course there are no functional trade-offs in the generating model. While the modular structure of the functional mapping is captured by loadings, the pattern of opposite signs effectively reflects the geometric constraint of orthogonal decomposition.

Table 1: Percentage of total variance (λ) and loadings for first three principal components of the simulated data generated by equation 4

	PC1	PC2	PC3
λ	0.53	0.16	0.11
P_1	-0.76	-0.33	-0.02
P_2	-0.77	-0.28	0.07

P_3	-0.76	-0.3	-0.05
P_4	-0.67	0.52	0.53
P_5	-0.67	0.52	-0.53

A second misconception arises from the belief that a pattern of opposite loadings on PC3 is evidence of a trade-off holding PC1 (interpreted at Individual Quality) constant. Even if PC3 were something real, say muscle fiber type composition, the net functional correlation between 1500 m speed and sprint speed is the sum of *all* the component correlations, as emphasized above. That is, this misconception lies in ignoring the loadings on the other principal components. Here is the full path model without the path coefficients



(5)

The math to compute expected correlations is exactly as before. The correlation between 1500 m speed and sprint speed is precisely equal to

$$\text{COR}(1500, \text{sprint}) = \sum \beta_{1500.j} \beta_{sprint.j} \tag{6}$$

where the β are the loadings and j indexes the jth principal component. If we want to say there is a trade-off between 1500 m speed and sprint speed holding PC1 ("Individual Quality") constant, we cannot just compute the correlations as a function of the loadings on PC3 but need the correlation due to all the factors except PC1, which is the sum in equation 6 minus that due to PC1. This correlation is equivalent to the *covariance* of the regression residuals in Method 3 above.

References

Aitchison, J. (2003). A concise guide to compositional data analysis. `http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf`.

Garland Jr, T. (1994). Quantitative genetics of locomotor behavior and physiology in a garter snake. In *Quantitative genetic studies of behavioral evolution* (ed. C. R. B. Boake), pp. 251–277. Chicago: University of Chicago Press.

Mitteroecker, P. and Bookstein, F. L. (2009). Examining modularity via partial correlations: A rejoinder to a comment by Paul Magwene. *Systematic Biology* p. syp040.

Van Damme, R., Wilson, R. S., Vanhooydonck, B. and Aerts, P. (2002). Performance constraints in decathletes. *Nature* **415**, 755–756.

Van Noordwijk, A. J. and De Jong, G. (1986). Acquisition and allocation of resources: their influence on variation in life history tactics. *American Naturalist* **128**, 137–142.

Walker, J. A. and Caddigan, S. P. (2015). Performance trade-offs and individual quality in decathletes. *Journal of Experimental Biology* .

Wilson, R. S., Niehaus, A. C., David, G., Hunter, A. and Smith, M. (2014). Does individual quality mask the detection of performance trade-offs? A test using analyses of human physical performance. *The Journal of Experimental Biology* **217**, 545–551.

Wright, S. (1918). On the nature of size factors. *Genetics* **3**, 367–374.

Wright, S. (1932). General, group and special size factors. *Genetics* **17**, 603–619.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics* .