In the discussion of the first simulation and Fig. 1, you seem to get a little ahead of yourself. You start comparing PC, PFC, and OLS well before you give us any reason to believe that OLS is a valid comparison. As I understand it, the comparison with OLS depends on two assumptions: that $\mathrm{E}(y|x)$ is linear in x and that Γ has only one column – as in the simulation model (7). I think a casual reader would not pick up either of those points (especially the latter) and in any case, they should be explained prior to discussing the relative merits of the three methods. Another point that does not seem clear to me is that I believe, but am not completely positive, that for this simulation PFC is the correct model, so that its superior performance should not be surprising.

I have never been a fan of partial least squares. In fact, I believe that only a bad estimation procedure used in the algorithm keeps it from being idiotic. Nonetheless, it is a procedure many people have used for dimension reduction in these problems and I was wondering if you wanted to mention it.

My first reaction to model (2) is that it was strikingly similar to the factor analysis model. Do you want to mention any relationship?

Page 5, last line: I believe "statistics" should be singular.

Page 8, line −10: I believe "setting" should be plural.

Page 12, line 1: "then, apart from constants,"

Page 13, last paragraph of Section 3: Please explain more.

Page 18: None of Section 5.2 was clear to me.

Page 21: I believe the reference to Cox (1972) should be to Cox (1968).

Page 20, paragraph around (10): I do not understand why there is a distinction to be made from what you were doing for model (2). Here you are minimizing $|G_0\hat{\Sigma}G_0|$ which, in the context of this discussion, does not seem very different to me from minimizing $\mathrm{tr}[\hat{\Sigma}Q_G] = \mathrm{tr}[G_0'\hat{\Sigma}G_0]$.

Page 22, first display: I believe what you are saying is that you can pick M_0 to maximize the first line of this display and pick M to maximize the second line, which gives you (13). Can you give a reference for that result? It certainly is not obvious to me.

Page 23, two lines above first display: I believe "subsets" should be singular.

I really liked the observation after Prop. 5.

Section 6.4: In Figure 2a,b is there any reason you do not look at variances less than 1? For both simulations, it might be good to explain why you chose to look at angles between estimates rather than some other measure. It is probably worth reiterating that, as in the first simulation, it is the nature of Γ that makes it appropriate to compare your results with those of a forward regression (y on x).

Page 25, line 4: I think it would be clearer to say, "set, and in the simulation when"

Page 25: In the discussion of Figure 2d, initially I wondered how much this has to do with looking at PFC(all) as opposed to just having a larger set of vectors to search over – but your discussion makes it clear that it matters where you get the vectors from.

However, I had a big problem with the fact that the residual candidate works well for $\sigma_0 \doteq 1$ because it seems to me that you should have the same problems there as you have with the PFC(PC) when $\sigma_0 \doteq \sqrt{2}$.

Page 28, line 10: I think it should be σ_0 not σ_Y.

Page 30, line 1: I think it should be "the fact that"

Page 31, line -2: "connection with OLS requireS"

Page 31, last line: This almost gives me the impression that you are thinking of these procedures as an end product. Throughout, I've been thinking of this as finding a good starting point for regressing y on x (see attached discussion). The discussion in the middle of page 36 seems to confirm the end product point of view, if that is so, it is something I totally missed in the paper – perhaps because I had already set my mind in a different direction. On the other hand, the discussion at the bottom of page 36 seems to confirm my view.

Page 32: On line 9, drop one "subspace". On line 14, "provide"

Figure 3: Isn't the moral of this simulation that (a) knowing Δ is such a huge advantage that it dwarfs all other issues and (b) when you don't know Δ, with this simulation where the true x on y relationship is a simple linear regression, modeling x on y using either a simple linear regression (OLS) or a cubic polynomial (PFC-Poly) works much better than using a step function with 8 levels (SIR)?

Page 34, third line above Discussion: "The same limitation do no occur" Second line of Discussion: "As a consequence"? Tenth line of Discussion: "from a recent dimension reduction methods"

Page 35, Fisher quotation: My 1970 version of the book has, "regarded as (i.) the study" – no second "as".

Page 39, first line of D: "it is known"

Page 43, Hotelling reference: "complex"

Dimension reduction, nonparametric regression, and multivariate linear models.

It seems to me that the key issue in this development is whether Cook's models (2), (5), (9), and (12) are broadly reasonable. The question did not seem to be extensively addressed. If they are appropriate, the results in the corresponding Propositions are rather stunning. It has long been known that the best regression model available – technically the best predictor of a random variable y based on a p dimensional random vector x – is the conditional mean $\mathrm{E}(y|x)$. The problem with this result is that it requires us to know the joint distribution of (x', y). Most of what we commonly recognize as regression analysis is an attempt to model the relationship $\mathrm{E}(y|x)$. This includes linear regression, nonlinear regression, generalized linear models, and the various approaches to "nonparametric" (actually, highly parametric) regression. Under the models being considered, there exists a $p \times d$ matrix Γ such that

$$y|x \sim y|\Gamma' x.$$

This means that $\mathrm{E}(y|x) = \mathrm{E}(y|\Gamma' x)$ regardless of what modeling strategy we choose to use. If anything, this dimensionality reduction from p to d is of more importance to nonparametric regression than other forms because, as the number of predictor variables increases, nonparametric regression gets hit harder by the curse of dimensionality than less highly parametric forms. As a result, nonparametric regression should benefit most from the existence of a generally valid reduction in dimensionality.

The issue with Cook's models is to estimate the column space of Γ, say, $C(\Gamma)$. In the first six sections, the results are all closely tied to the eigenvectors (principal component vectors) of some estimated covariance matrix for the predictor variables x, say $\hat{\Sigma}$. For model (2), the space is spanned by the first d principal component vectors of the usual $\hat{\Sigma}$. For model (5), the space is spanned by the first d principal component vectors of a restricted version of $\hat{\Sigma}$. For models (9) and (12), the estimation procedure is a bit more complicated. The key is

that for both models (9) and (12) the population covariance matrix of x can be written as

$$\Sigma = \Gamma V D V' \Gamma + \Gamma_0 V_0 D_0 V_0' \Gamma_0$$

with D and D_0 diagonal matrices in such a way that

$$\Sigma(\Gamma V) = (\Gamma V) D, \qquad \Sigma(\Gamma_0 V_0) = (\Gamma_0 V_0) D_0.$$

This implies that the eigenvectors of Σ are either in $C(\Gamma)$ or in $C(\Gamma_0) \equiv C(\Gamma)^{\perp}$, the orthogonal complement of $C(\Gamma)$. The problem is to figure out which d out of the p orthogonal eigenvectors belong in $C(\Gamma)$. To estimate $C(\Gamma)$, find the orthogonal eigenvectors of $\hat{\Sigma}$, say, $v_1, \ldots, v_p$ and check the likelihood of every one of the p choose d combinations that has d of the v_is in $C(\Gamma)$ and the remaining $p - d$ vectors in the orthogonal complement. Whichever combination maximizes the likelihood, provides the estimate of $C(\Gamma)$. In case $\binom{p}{d}$ is large, Cook provides a sequential selection method. The key difference between the procedures for models (9) and (12) is that the likelihoods are different.

The remainder of my discussion is an attempt to put the question of estimating the reduced space into the context of multivariate linear model theory. To do this, I will change Cook's notation completely, so that the problem looks more like standard multivariate linear models, but then re-identify the parts of the problem that interest Cook. I do not presume that any of this is new to Cook, but I found it helpful in understanding the process.

In discussing multivariate linear models, liberal use is made of Kronecker products, Vec operators, and their properties, see, for example, Christensen (2002, Definition B.5 and Subsection B.5). Recall also that for a univariate linear model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = V$,

$$SSE \equiv (Y - X\hat{\beta})' V^{-1} (Y - X\hat{\beta}) = Y' V^{-1} (Y - X\hat{\beta}).$$

Moreover, least squares estimates are BLUEs, and thus maximum likelihood, if $C(VX) \subset C(X)$. We will apply these facts to the multivariate models. Finally, let J_r^s denote an $r \times s$

matrix of 1s with $J_r \equiv J_r^1$ and let $P_A = A(A'A)^- A'$ be the perpendicular projection operator (ppo) onto $C(A)$ with $\mathrm{r}(A)$ the rank of A.

The standard multivariate linear model involves dependent variables $y_1, \ldots, y_q$. If n observations are taken on each dependent variable, we have y_{ih}, $i = 1, \ldots, n$, $h = 1, \ldots, q$. Let $Y_h = [y_{1h}, \ldots, y_{nh}]'$ and $y_i' = [y_{i1}, \ldots, y_{iq}]$. For each h, we have a linear model.

$$Y_h = X\beta_h + e_h, \quad \mathrm{E}(e_h) = 0, \quad \mathrm{Cov}(e_h) = \sigma_{hh} I,$$

where X is a known $n \times p$ matrix that is the same for all dependent variables, but β_h and the error vector $e_h = [e_{1h}, \ldots, e_{nh}]'$ are peculiar to the dependent variable Y_h.

The multivariate linear model consists of fitting the q linear models simultaneously. Letting

$$Y_{n\times q} = [Y_1, \ldots, Y_q], \quad B_{p\times q} = [\beta_1, \ldots, \beta_q], \quad e_{n\times q} = [e_1, \ldots, e_q],$$

the multivariate linear model is

$$Y = XB + e\,. \tag{1}$$

Alternatively, thinking of X as a matrix with rows x_i' and e as having rows ε_i', we can write the multivariate linear model as

$$y_i' = x_i' B + \varepsilon_i', \quad i = 1, \ldots, n.$$

To perform maximum likelihood, we assume that the ε_is are independent $N(0, \Sigma)$ random vectors. It is reasonably well known that for any ppo P_A,

$$\mathrm{E}_{Y|X}(Y'P_A Y) = \mathrm{r}(A)\Sigma + B'X'P_A XB. \tag{2}$$

Σ is now being used for the conditional covariance matrix of $y|x$, whereas Cook used Σ for the marginal covariance matrix of x.

A generalization of the multivariate linear model that is often associated with growth curve models, cf. Christensen (2001) is

$$Y = X\Gamma Z' + e \tag{3}$$

where the unknown parameter matrix B in (1) is replaced by the product of a reduced parameter matrix Γ that is $p \times d$ and a fixed, known matrix Z that is $q \times d$ with $\mathrm{r}(Z) \leq d < q$. This is essentially Cook's model (5) when applied to data and using drastically different notation. (Our Y is his X, our X is his known function of y, F_y, our Z is his Γ. etc.) The ultimate goal of our exercise is to drop the assumption that we know Z and estimate it, or more properly $C(Z)$, from the data. But for now, we act as if Z is known. Note that the "growth curve" model specifies something akin to a linear model for each row of Y,

$$y_i = Z(\Gamma x_i) + \varepsilon_i. \quad i = 1, \ldots, n$$

Moreover, using Kronecker products and Vec operators, we can turn the multivariate growth curve model (3) into a univariate linear model.

$$\mathrm{Vec}(Y) = [Z \otimes X]\mathrm{Vec}(\Gamma) + \mathrm{Vec}(e)\,. \tag{4}$$

with

$$\mathrm{Vec}(e) \sim N(0, [\Sigma \otimes I_n]).$$

There are a couple of refinements to model (4) used in Cook's development. First, the growth curve model is specified as

$$Y = J\mu' + X\Gamma Z' + e\,. \tag{5}$$

with $J'X = 0$ and $Z'Z = I_d$. As a linear model (5) becomes

$$\mathrm{Vec}(Y) = [I_q \otimes J_n]\mu + [Z \otimes X]\mathrm{Vec}(\Gamma) + \mathrm{Vec}(e)\,.$$

Second is the assumption in Cook's models (2) and (5) that

$$\Sigma = \sigma^2 I_q,$$

in which case,

$$\mathrm{Cov}[\mathrm{Vec}(e)] = \sigma^2[I_q \otimes I_n] = \sigma^2 I_{nq},$$

so standard estimation results apply to the model. In particular, least squares estimates of the parameters μ and $\mathrm{Vec}(\Gamma)$ are maximum likelihood estimates and the likelihood function for fixed σ^2 evaluated at the maximum likelihood estimates of μ and Γ is, ignoring the constant,

$$-\frac{nq}{2}\log(\sigma^2) - \frac{SSE}{2\sigma^2}. \tag{6}$$

Performing the usual computations necessary to finding least squares estimates, but using properties of Kronecker products and Vec operators and exploiting the fact that since $J'X = 0$ we have $C([I_q \otimes J_n]) \perp C([Z \otimes X])$ so that estimation of μ and Γ can be performed separately, the least squares estimates reduce to

$$\hat{\mu} = \bar{y}_\cdot \qquad \hat{\Gamma} = (X'X)^- X'YZ(Z'Z)^-$$

or, alternatively,

$$X\hat{\Gamma}Z = P_X Y P_Z.$$

The maximum likelihood estimate of σ^2 is obtained by differentiating (6) with respect to σ^2 and setting it equal to 0, yielding $\hat{\sigma}^2 = SSE/nq$. In particular, the perpendicular projection operator onto $C([I_q \otimes J_n], [Z \otimes X])$ is $[I_q \otimes (1/n)J_n^n] + [P_Z \otimes P_X]$, so

$$\begin{aligned} SSE &= \mathrm{Vec}(Y)'[I_q \otimes (I - (1/n)J_n^n)]\mathrm{Vec}(Y) - \mathrm{Vec}(Y)'[P_Z \otimes P_X]\mathrm{Vec}(Y) \\ &= \|\mathrm{Vec}[(I - (1/n)J_n^n)Y]\|^2 - \|\mathrm{Vec}(P_X Y P_Z)\|^2 \\ &= \mathrm{tr}[Y'(I - (1/n)J_n^n)Y] - \mathrm{tr}[P_Z Y' P_X Y P_Z]. \end{aligned} \tag{7}$$

Using notation analogous to Cook's, three estimators that we will use frequently are

$$\hat{\Sigma} \equiv \frac{1}{n}Y'\left(I - \frac{1}{n}J_n^n\right)Y, \quad \hat{\Sigma}_{\text{fit}} \equiv \frac{1}{n}Y'P_X Y, \quad \hat{\Sigma}_{\text{res}} \equiv \frac{1}{n}Y'\left(I - \frac{1}{n}J_n^n - P_X\right)Y.$$

Note that $\hat{\Sigma}$ is the maximum likelihood estimate of the covariance matrix when fitting the usual multivariate one-sample model $Y = J_n\mu' + e$. Using the assumption, $Z'Z = I_d$ so that $ZZ' = P_Z$,

$$SSE = \mathrm{tr}[n\hat{\Sigma}] - \mathrm{tr}[Z'n\hat{\Sigma}_{\text{fit}}Z].$$

We are finally in a position to address Cook's question, the fact that we do not actually know Z. To maximize the likelihood (6) as a function of Z we need to maximize $\mathrm{tr}[Z'n\hat{\Sigma}_{\mathrm{fit}}Z]$ as a function of Z subject to $Z'Z = I_d$. If we think about finding the columns of Z sequentially, i.e., finding z_1 to maximize $z'\hat{\Sigma}_{\mathrm{fit}}z$ subject to $\|z_1\|^2 = 1$, then finding z_2 to maximize $z'\hat{\Sigma}_{\mathrm{fit}}z$ subject to $\|z_2\|^2 = 1$ and $z_2'z_1 = 0$, and so on, this is a standard problem in multivariate analysis that is solved by finding the eigenvectors of $\hat{\Sigma}_{\mathrm{fit}}$ relative to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$. Of course, since Z is $q \times d$, we consider only the first d eigenvectors.

To examine a model equivalent to Cook's model (2), we consider the most extreme (largest) choice for X which is $C(X) = C(J_n)^{\perp}$. In this case, $P_X = I_n - (1/n)J_n^n$ so that $\hat{\Sigma}_{\mathrm{fit}} = \hat{\Sigma}$. It follows that the maximum likelihood estimate of Z consists of the first d principal component vectors. As pointed out by Cook, the number of parameters in our Γ matrix is pd. However, with this choice of X, $p = n - 1$, so the number of parameters rises with the sample size.

For Cook's models (9) and (12) in section 6, the covariance structure changes. As indicated earlier, the estimation methods ultimately involve determining which of the principal component directions are most likely where principal components can be computed from some estimate of Σ, which may be any, or preferably all, of $\hat{\Sigma}$, $\hat{\Sigma}_{\mathrm{fit}}$, or $\hat{\Sigma}_{\mathrm{res}}$. For Cook's model (12) we again have

$$\mathrm{Vec}(Y) = [I_q \otimes J_n]\mu + [Z \otimes X]\mathrm{Vec}(\Gamma) + \mathrm{Vec}(e)$$

but recalling that $Z'Z = I_d$, we now incorporate a matrix Z_0 with $Z_0'Z = 0$ and $Z_0'Z_0 = I_{q-d}$ and assume

$$\mathrm{Vec}(e) \sim N(0, [Z_0\Omega_0^2Z_0' + Z\Omega^2Z' \otimes I_n]).$$

Observe that least squares estimates will still be BLUEs and thus maximum likelihood estimates because $C([Z_0\Omega_0^2Z_0' + Z\Omega^2Z' \otimes I_n][Z \otimes X]) \subset C([Z \otimes X])$.

The SSE now involves the perpendicular projection operators as in (7) but also involves the inverse of the covariance matrix. With our assumptions about Z and Z_0,

$$[Z_0\Omega_0^2 Z_0' + Z\Omega^2 Z' \otimes I_n]^{-1} = [Z_0\Omega_0^{-2} Z_0' + Z\Omega^{-2} Z' \otimes I_n].$$

The SSE becomes

SSE

$$\begin{aligned}
&= \mathrm{Vec}(Y)'[(Z\Omega^{-2}Z' + Z_0\Omega_0^{-2}Z_0') \otimes (I - (1/n)J_n^n)]\mathrm{Vec}(Y) - \mathrm{Vec}(Y)'[Z\Omega^{-2}Z' \otimes P_X]\mathrm{Vec}(Y) \\
&= \mathrm{Vec}(Y)'\mathrm{Vec}[(I - (1/n)J_n^n)Y(Z\Omega^{-2}Z' + Z_0\Omega_0^{-2}Z_0')] - \mathrm{Vec}(Y)'\mathrm{Vec}(P_X Y Z\Omega^{-2}Z') \\
&= \mathrm{tr}[Y'(I - (1/n)J_n^n)Y(Z\Omega^{-2}Z' + Z_0\Omega_0^{-2}Z_0')] - \mathrm{tr}(Y'P_X Y Z\Omega^{-2}Z') \\
&= \mathrm{tr}[\Omega_0^{-2}Z_0'Y'(I - (1/n)J_n^n)YZ_0] + \mathrm{tr}\{\Omega^{-2}Z'[Y'(I - (1/n)J_n^n)Y - Y'P_X Y]Z\} \\
&= \mathrm{tr}[\Omega_0^{-2}Z_0' n\hat{\Sigma} Z_0] + \mathrm{tr}\{\Omega^{-2}Z'[n\hat{\Sigma} - n\hat{\Sigma}_{\mathrm{fit}}]Z\}
\end{aligned}$$

The likelihood will be

$$\begin{aligned}
\frac{-n}{2}\log(|Z_0\Omega_0^2 Z_0' + Z\Omega^2 Z'|) + \frac{-1}{2}SSE &= \frac{-n}{2}\log(|\Omega_0^2|) + \frac{-n}{2}\mathrm{tr}[\Omega_0^{-2}Z_0' n\hat{\Sigma} Z_0] \\
&\quad + \frac{-n}{2}\log(|\Omega^2|) + \frac{-n}{2}\mathrm{tr}[\Omega^{-2}Z'[n\hat{\Sigma} - n\hat{\Sigma}_{\mathrm{fit}}]Z]
\end{aligned}$$

which, maximizing over Ω and Ω_0, Cook indicates reduces to

$$\frac{-n}{2}\log(|Z_0' n\hat{\Sigma} Z_0|) + \frac{-n}{2}\log(|\Omega^{-2}Z'[n\hat{\Sigma} - n\hat{\Sigma}_{\mathrm{fit}}]Z|).$$

As before, Cook's model (9) can be viewed as the special case where $C(X) = C(J_n)^{\perp}$, so that the second term in the likelihood disappears.

To continue this discussion, we need to leave the conditional world of linear models and consider the unconditional expected values of $\hat{\Sigma}$, $\hat{\Sigma}_{\mathrm{fit}}$, and $\hat{\Sigma}_{\mathrm{res}}$. Conditionally, applying equation (2) to model (5) when $J'A$=0 gives

$$\mathrm{E}_{Y|X}(Y'P_A Y) = \mathrm{r}(A)\,\Sigma + Z\Gamma' X' P_A X \Gamma Z'. \tag{8}$$

For $\hat{\Sigma}$ and $\hat{\Sigma}_{\text{fit}}$ the appropriate ppo has $P_A X = X$ and for $\hat{\Sigma}_{\text{res}}$ the ppo has $P_A X = 0$. We have assumed that $J'X = 0$ which is only reasonable if the random rows of X have been adjusted by their sample means, nonetheless it is reasonable to define the marginal covariance matrix of a row of X as

$$V_x \equiv \frac{1}{n-1}\mathrm{E}_X(X'X).$$

These results quickly yield the following expectations

$$\mathrm{E}(\hat{\Sigma}) = \frac{n-1}{n}\Sigma + \frac{n-1}{n}Z\Gamma' V_x \Gamma Z',$$

$$\mathrm{E}(\hat{\Sigma}_{\text{fit}}) = \frac{\mathrm{r}(X)}{n}\Sigma + \frac{n-1}{n}Z\Gamma' V_x \Gamma Z',$$

$$\mathrm{E}(\hat{\Sigma}_{\text{res}}) = \frac{n-1-\mathrm{r}(X)}{n}\Sigma.$$

In particular, with $\Sigma = Z_0\Omega_0^2 Z_0' + Z\Omega^2 Z'$, Cook's Proposition 4 says that the estimates converge in probability to the limits of their expected values.

Cook's second simulation has a true model with $d = 1$, $n = 250$, $q = 10$ (his p), $p = 1$, $\Gamma = 1$, $\Omega = \sigma$, $\Omega_0 = \sigma_0 I_9$, and $V_x = \sigma_x^2$ (his σ_Y^2). Here,

$$\mathrm{E}(\hat{\Sigma}) = \frac{n-1}{n}\sigma_0^2 Z_0 Z_0' + \frac{n-1}{n}(\sigma^2 + \sigma_x^2)ZZ',$$

$$\mathrm{E}(\hat{\Sigma}_{\text{fit}}) = \frac{\mathrm{r}(X)}{n}\sigma_0^2 Z_0 Z_0' + \left(\frac{\mathrm{r}(X)}{n}\sigma^2 + \frac{n-1}{n}\sigma_x^2\right)ZZ',$$

$$\mathrm{E}(\hat{\Sigma}_{\text{res}}) = \frac{n-1-\mathrm{r}(X)}{n}\sigma_0^2 Z_0 Z_0' + \frac{n-1-\mathrm{r}(X)}{n}\sigma^2 ZZ'.$$

Cook's simulation results make good sense in terms of these expected values. Terms involving $\mathrm{r}(X)/n$ should be unimportant. When σ_0 is small, $\mathrm{E}(\hat{\Sigma})$ is dominated by $(\sigma^2+\sigma_x^2)ZZ'$ which is larger than the corresponding terms $\sigma_x^2 ZZ'$ and $\sigma^2 ZZ'$ for $\hat{\Sigma}_{\text{fit}}$ and $\hat{\Sigma}_{\text{res}}$, respectively, so it works best. When σ_0 is comparable to σ and σ_x, $\hat{\Sigma}_{\text{fit}}$ works well, because $\mathrm{E}(\hat{\Sigma}_{\text{fit}})$ is much less affected by σ_0^2 than the other estimates. And when σ_0 is large, $\hat{\Sigma}_{\text{res}}$ works well because then we need to look at the eigenvectors for small eigenvalues of $\hat{\Sigma}_{\text{res}}$ and $\hat{\Sigma}$, but the term

$\sigma^2 ZZ'$ for $\hat{\Sigma}_{\text{res}}$ is smaller than the term $(\sigma^2 + \sigma_x^2)ZZ'$ for $\hat{\Sigma}$, whereas the expected value of $\hat{\Sigma}_{\text{fit}}$ is relatively unaffected by σ_0^2 getting large. As Cook mentions, when $\sigma_x^2 + \sigma^2 = \sigma_0^2$, there is very little ability for $\hat{\Sigma}$ to identify $C(Z)$ because then $\mathrm{E}(\hat{\Sigma}) = (n-1)\sigma_0^2/n\, I_q$, so we cannot really expect the eigenvectors of $\hat{\Sigma}$ to help us identify $C(Z)$. Similarly, when $\sigma^2 = \sigma_0^2$, $\mathrm{E}(\hat{\Sigma}_{\text{res}}) = (n - 1 - \mathrm{r}(X))\sigma_0^2/n\, I_q$.

The expectations of the estimates also show that there should almost always be some ability to estimate $C(Z)$, cf. Cook's section 7. In particular,

$$\mathrm{E}\left[\frac{n}{\mathrm{r}(X)}\hat{\Sigma}_{\text{fit}} - \frac{n}{n-1}\hat{\Sigma}\right] = \frac{n-1-\mathrm{r}(X)}{\mathrm{r}(X)} Z\Gamma' V_x \Gamma Z',$$

so the first d principal component vectors of the estimate $\frac{n}{\mathrm{r}(X)}\hat{\Sigma}_{\text{fit}} - \frac{n}{n-1}\hat{\Sigma}$ should be at least a reasonable estimate of a basis for $C(\Gamma)$. For large samples this is similar to looking at the directions determined by $\hat{\Sigma}_{\text{fit}}$ but in the extreme case of $C(X) = C(J_n)^{\perp}$, the estimator is degenerate at 0. Of course, according to Cook's Proposition 7, for general Σ (Cook's $\sigma^2\Delta$), it no longer suffices to estimate $C(Z)$, we need to estimate $C(\Sigma^{-1}Z)$. Fortunately, $\hat{\Sigma}_{\text{res}}$ provides an estimate of Σ, so we can just transform the estimated basis for $C(\Gamma)$. For large n, it makes sense to base estimation of $C(\Gamma)$ on $\hat{\Sigma}_{\text{fit}}$, but rather than transforming its eigenvectors, one could alternatively look directly at the eigenvectors of $\hat{\Sigma}_{\text{res}}^{-1}\hat{\Sigma}_{\text{fit}}$, which is Cook's recommendation when $p = d$. This intuitive approach based on the expected values of (matrix) quadratic forms seems analogous to using Henderson's method three for estimating variance components, whereas Cook is recommending a maximum likelihood procedure, which I suspect is better.

I found Cook's discussion of standardization in Section 7.3 disturbing. I am not dogmatic about the need to standardize variables prior to finding principal components. When the measurements are all on similar scales, using the original scales seems reasonable to me, as when measuring the height, length, and width of turtle shells in centimeters. On the other hand, if I measure length in kilometers and height and width in millimeters, the first principal component will essentially ignore the lengths, regardless of any role that length might play

in prediction. I suspect that the point of Cook's discussion is that in a situation where you need to standardize the variables, there will be little reason to suppose that his models (2) or (5) are appropriate, which means there is little reason to use principal components.

As I indicated at the beginning of my discussion, my biggest problem with these procedures is that I do not have a good feel for when the various models will be appropriate. Multivariate linear model theory should allow us to use $\hat{\Sigma}_{\text{res}}$ to test the assumption of Cook's models (2) and (5) that $\Sigma = \sigma^2 I$. I am less sure if it will provide a test of whether $\Sigma = \sigma^2 ZZ' + \sigma_0^2 Z_0 Z_0'$, when Z is unknown, but a generalized likelihood ratio test seems plausible. In any case, the procedures in Section 7.2 seem generally applicable.

REFERENCES

Christensen, Ronald (2001). *Advanced Linear Modeling*, Second Edition. Springer, New York.

Christensen, Ronald (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, Third Edition. Springer, New York.