PCA with missing data

1. I’m now an adjunct professor at UVic, an unpaid employee of the University but without obvious duties. The benefit on my side is the access to the library, attendance to lectures and colloquia, and email facility.
2. Today I’d like to talk about missing data in PCA.
3. But before we jump into the topic, I’d like to briefly introduce projectors that appear throughout my talk. Here is the orthogonal projector onto Sp(Z), and its complement. And this is the projector onto Sp(Z) along Ker(Z’W) for an nnd metric matrix W that satisfy this rank condition.
4. I need several other symbols to begin my talk:
   X = […, x_j, …]
   1_n
   D_wj
   F
   A
   u_j, u_0j
   x*_j
5. There have been at least two approaches to missing data in PCA. One is based on HA, and the other on WLRA. In HA we set up a criterion like this to be minimized. That is, the data vectors and the constant term are weighted such a way that the resultant matrices are all as close as possible to the matrix o component scores. Here, the indicator matrices D_wj are used to weigh the element of the SS terms. In the original formulation of the problem by Meulman (1982), the columnwise centered vector x*_j was used instead of the raw data vector x_j and the constant terms were forced to be zero. We presented here a little more general formulation in which we use x_j and include nonzero constant terms.
6. The second approach, on the other hand, set up a criterion like this, called the WLRA. Here we approximate the data X* by a low rank matrix. This criterion was first proposed by Gabriel and Zamir (1979). We could have used the raw data vector x_j here and include the constant term in line with the HA criterion. We will discuss this extension later on. However, it will introduce some complication in the algorithm used to estimate parameters.
7. Three remarks: (a) The criteria similar in form were originally proposed by Meredith (1964) in the context of factor matching. (b) They are simply related when the data are complete. They are, however, distinct when there are missing data. (c) HA leads to closed-form solutions, while WLRA iterative solutions.
8. We start with HA. There are two alternative solutions to the HA problem that I am aware: One is called the MDP solution originally proposed for multiple correspondence analysis with missing data. The other is called the TE method proposed by Shibayama (1988) in the context of test equation. In university entrance examinations in Japan, not all applicants take exactly the same examinations, some English as the second foreign language, others French, for example. This creates massive missing data, yet the test administrator has to come up with a set of scores which can rank order all applicants for admission. These two methods of solutions were invented in completely different contexts, and were thought to be distinct. This turned out to be false in 2003.

9. To show their essential equivalence, we first rewrite the HA criterion without using the summation mark. This simplifies the solution, although not necessarily computationally more efficient. Let
   J_n: p identity matrices of order n placed on top of each other
   D_1: a variable indicator marix
   D_X: a block diagonal matrix of x_j
   D_w: a block diagonal matrix of D_wj
   U' and U_0': matrices of u_j' and u_0j
   Then, \phi can be rewritten in this form.
10. The MDP method minimizes \phi in this order: First, wrt U_0 conditional on U and F, then wrt U conditional on F, and finally wrt to F under these restrictions on F. These restrictions are necessary because \phi can be trivially made zero by setting all parameter matrices to zero matrices, and J_n and D_1 are not disjoint. Anyway, this leads to the GEQ of this form, where \tilde{Q} is …
11. In the TE method, the same criterion is minimized in a different order, which entails different restrictions under which \phi is minimized, which in turn leads to different parameter values. So let us rewrite the criterion using different symbols for parameter matrices (G for F, and V and V_0 for U ad u-0, respectively), and call this criterion \psi,
12. which is minimized first wrt G conditional on V nd V_0, then wrt V_0 conditinal on V, and finally wrt V under these restrictions. This leads to this GEQ of this form. This looks different from the one derived previously.
13. However, the two GEQ are simply related. Let X* be defined this way, and note that Q defined previously has this expression, and that this expression in the second GEQ can be rewritten as this due to a lemma given by Takane and Zhou (2012). Then, the two GEQ can be rewritten as

    respectively. Now the relationship is obvious. This is like the EQ for A'A and this AA', which are simply related via GSVD of this matrix with metric matrices Q and S.
14. Perhaps the most economical way of calculating the parameter estimates would be to use the second GEQ to obtain V from which F and U can be easily calculated, and finally U_0.
15. Now we are going to talk about the second approach to missing data in PCA, namely the one based on WLRA, which allows low rank approximations of data matrices under an extremely flexible weighting scheme. Let x* = vec(X*), and the vectorized version of a low rank approximation to X*, denoted by x_0 = vec(X_0). Remember that here we use the columnwise centred data matrix X*. We minimize this criterion wrt x_0, where $W* is an np by np matrix of weights, which can be any symmetric pd matrix.

    This criterion cannot be minimized in closed form except in the special case in which W* can be factored into the Kronecker product of two pd matrices, in which case it reduces to GSVD(X*)_{K,L}.
16. For a fixed rank r, X_0 can be reparameterized as FA', so that x_0 has two alternative expressions, x_0 = (A \otimes I_n) vec(F) = (I_p \otimes F)\vec(A'). We may use these expressions to develop a monotonically convergent iterative algorithm to update F and A alternately.

17. This is the most general algorithm, which can be simplified considerably when W* is diagonal. In this case, \tau can be rewritten two alternative ways, one given here and the other on the next slide. This indicates a_j can be updated separately for each j, and
18. this indicates f_i can be updated separately for each i. Again we iterate between updatings of A and F until convergence.
19. So far in LRA, we have assumed that we have the columnwise centered data matrix. In this case the centring is performed wrt non-missing observations, which is fine if missing data occur randomly. In this case the mean calculated based on observed data are fairly close to those that would have been obtained if the data re complete. This may not be so in entrance examinations because applicants choose to take exams they think they are good at, and indeed this was the basic motivation to use the raw data matrix and the constant terms in the TE method. What if we use the raw data and include the constant term in the WLRA method. In this case we minimize this criterion wrt to F, A and m, which leads to the updates of m and a given here, where Q_j is given by …. This Q_j matrix is not diagonal, and when \tau* is minimize wrt F, it is no longer separable wrt i, which entails a full version of the algorithm in this phase of updating.
20. I tend to prefer the method based HA on a priori grounds,

    although systematic Monte Carlo studies are essential to compare te two techniques.