Orthogonal, but not Orthonormal, Procrustes Problems

Richard Everson
Laboratory for Applied Mathematics,
CUNY/Mount Sinai,
New York, NY, 10029
rme@camelot.mssm.edu

March 12, 1997

Abstract

The classical matrix Procrustes problem seeks an orthogonal matrix, U, which most closely transforms a given matrix into a second matrix. We consider the Procrustes problem in which the requirement that the columns of U be orthonormal is relaxed to orthogonality. Closed form solutions cannot be found, but numerical schemes to find the best matrix (in the Frobenius norm) are advanced. Numerical examples are given and the of the orthogonal Procrustes matrix alternative to Löwdin orthogonalization is discussed.

1 Introduction

The villain Procrustes forced his victims to sleep on an iron bed; if they did not fit the bed he cut off or stretched their limbs to make them fit. The classical Procrustes matrix problem (Horn and Johnson 1985; Golub and Loan 1983) asks how closely a matrix $A \in \mathbb{R}^{m\times n}$ can be approximated by a second, given, matrix $B \in \mathbb{R}^{p\times n}$ multiplied by a matrix $U \in \mathbb{R}^{m\times p}$ with orthogonal columns. Using the Frobenius norm, we therefore have the following problem:

$$\text{minimize } \|A - UB\|^2 \qquad \text{with } U^{\mathrm{T}}U = I_p \tag{1}$$

Since U has orthonormal columns,

$$\|A - UB\|^2 = \mathrm{Tr}A^{\mathrm{T}}A - 2\mathrm{Tr}AB^{\mathrm{T}}U^{\mathrm{T}} + \mathrm{Tr}B^{\mathrm{T}}B \tag{2}$$

Consequently (1) is equivalent to the problem of maximizing $\mathrm{Tr}AB^{\mathrm{T}}U^{\mathrm{T}}$. The maximizing U may be found in terms of the singular value decomposition (SVD) of AB^{T}. If

$$AB^{\mathrm{T}} = W\Sigma Z^{\mathrm{T}}, \tag{3}$$

where $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_p)$, is the SVD of AB^{T} we have

$$\mathrm{Tr}AB^{\mathrm{T}}U^{\mathrm{T}} = \mathrm{Tr}W\Sigma Z^{\mathrm{T}}U^{\mathrm{T}} = \mathrm{Tr}W^{\mathrm{T}}UZ\Sigma = \mathrm{Tr}\ T\Sigma = \sum_{i=1}^{p} t_{ii}\sigma_i \leq \sum_{i=1}^{p}\sigma_i \tag{4}$$

where $T = W^{\mathrm{T}}UZ$ is an orthogonal matrix. The trace is maximized when $T = I_p$ so that $U = WZ^{\mathrm{T}}$, which may be recognized as the orthogonal polar factor of AB^{T}.

The orthogonal Procrustes problem might properly be called the orthonormal Procrustes problem since the condition imposed on U is that it's columns be orthonormal. In this note we discuss the Procrustes problem in which the demand of orthonormality is relaxed to orthogonality. If $D \in \mathbb{R}^{p\times p}$ is a diagonal matrix we consider the following problem:

$$\text{minimize } \|A - UB\|^2 \qquad \text{with } U^{\mathrm{T}}U = D^2. \tag{5}$$

The companion problem, which is no longer equivalent to (5), is stated in term of different $A \in \mathbb{R}^{m\times p}$, $B \in \mathbb{R}^{m\times n}$ and $U \in \mathbb{R}^{n\times p}$,

$$\text{minimize } \|A - BU\|^2 \qquad \text{with} \qquad U^{\mathrm{T}}U = D^2. \tag{6}$$

In each case, we do not assume that D is *a priori* known. Our victims must sleep with their limbs held orthogonal, but they will not be amputated or stretched if they don't match the bed's dimensions.

The classical matrix problem finds applications in factor analysis and statistics (Gower 1984; Green 1952), in structural identification (Beattie and Smith 1992), in robotics and, when $B \equiv I$, in the re-orthogonalization of a basis. Watson (Watson 1994) gives numerical schemes to solve the orthogonal Procrustes problem using the Schatten p-norms. The current study is motivated by the analysis data derived from the optical imaging of the visual cortex (Everson et al. 1997).

2 $A - UB$

2.1 Tandem algorithm

The easier problem is (5) and we address this first. If D is known the substitution $U = VD$, where $V^{\mathrm{T}}V = I_p$ reduces (5) to the classical problem:

$$\text{minimize } \rho = \|A - VDB\|^2 \qquad \text{with} \qquad V^{\mathrm{T}}V = I_p, \tag{7}$$

and V is the found in terms of the orthogonal polar factor of $AB^{\mathrm{T}}D$, which we denote by $\mathrm{OPF}(AB^{\mathrm{T}}D)$ On the other hand, if D is unknown, but V is known, a solution may be found by differentiating ρ with respect to the vector $\mathbf{d} = \mathrm{diag}(D)$:

$$\frac{\partial \rho}{\partial d_k} = -2\sum_{i=1}^{m} \left[AB^{\mathrm{T}}\right]_{ik} V_{ik} + 2d_k \sum_{i=1}^{n} B_{ki}^2. \tag{8}$$

Setting the left hand side to zero yields d_k and we note that there is no need to check the uniqueness or character of the turning point since ρ is quadratic in d.

These two minimizations may be iterated in tandem to solve the full problem (5). Thus we have the following iteration scheme:

$$\mathbf{d}_1 = (1, ..., 1)^{\mathrm{T}} \tag{9a}$$

Repeat until converged, $i = 1, 2, 3...$ (9b)

Obtain (classical) $V_i = \mathrm{OPF}(AB^{\mathrm{T}}D_i)$ (9c)

Obtain (eq. (8)) $\mathbf{d}_{i+1}$ by minimizing $\|A - V_i D_{i+1} B\|^2$ (9d)

Go to (9b). (9e)

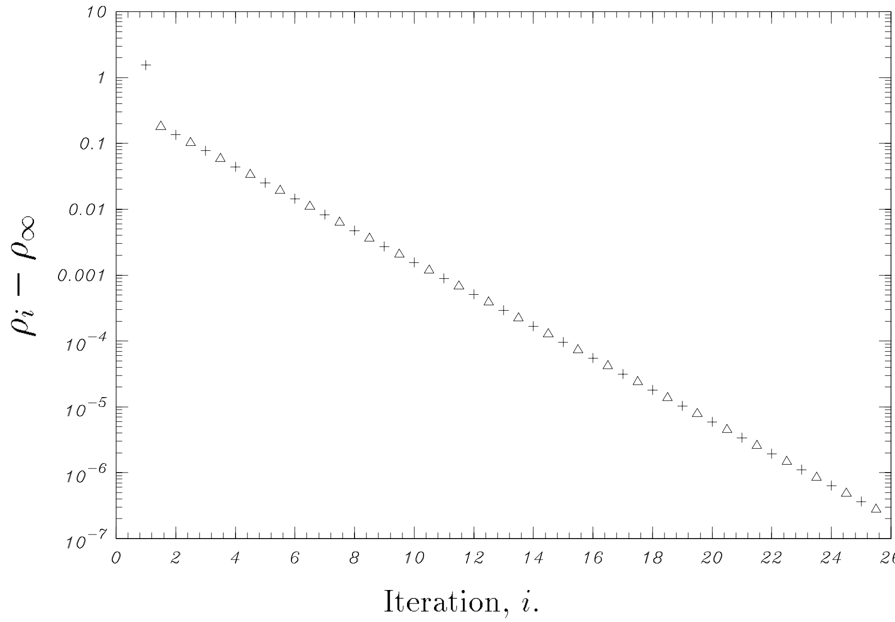

Figure 1: *Convergence of the tandem iteration (9). The differences between ρ at each stage of the iteration and ρ for the fully converged solution are plotted as crosses (9c) and triangles (9d).*

Since at each stage ρ is decreased the scheme converges, though not necessarily to the global minimum. In fact, convergence is linear. This is illustrated in Fig. 1 for matrices $A \in \mathbb{R}^{10\times 5}$ and $B \in \mathbb{R}^{3\times 5}$ whose entries are random numbers drawn from a uniform distribution between 0 and 1. The difference between ρ at each stage of the iteration and ρ for the solution converged to machine precision ($\approx 10^{-16}$) is plotted. Crosses at integer abscissae indicate ρ from (9c) and triangles at half-integer abscissae represent ρ from (9d). The values of ρ corresponding to the classical and fully converged U are 5.16 and 3.61, and the final $\mathbf{d} = (1.91, 1.24, 0.94)$.

We remark that at every stage of the iteration D is a diagonal matrix and V has orthonormal columns, so that each approximation to the solution lies on the manifold of admissible solutions. Although (9c) finds V_i as the orthogonal polar factor of $AB^{\mathrm{T}}B$, which at first sight requires the (expensive) SVD of $AB^{\mathrm{T}}D$, we mention that efficient iterative methods to calculate the orthogonal polar factor are available (Higham 1986).

2.2 Conjugate gradient

Instead of (9), ρ may also be minimized by a more direct numerical scheme in which the conjugate gradient method is used to optimize D and V simultaneously.

We wish to consider all candidate matrices $V \in \mathbb{R}^{m\times p}$ with orthonormal columns. Let V_0 be the solution to the classical Procrustes problem for A and B, that is, the orthogonal polar factor of AB^{T}. Then any V may be expressed as

$$V = V_0 G, \tag{10}$$

where $G \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. Clearly G has determinant ± 1. Further, since $G = V_0^{\mathrm{T}} V$ and $\det(G)$ is a continuous function of G, the determinant of G for the V that minimizes ρ must be the same as $\det(V_0^{\mathrm{T}} V_0) = +1$.

Since G is orthogonal with determinant $+1$, it may be parameterized by $p(p-1)/2$ independent quantities or angles (the analogues of Euler angles). Several different parameterizations are possible;

we choose to represent G as a product of $p(p-1)/2$ Givens or Jacobi elementary plane rotation matrices, which are matrices of the form

$$G(i,j,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & \sin(\theta) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \tag{11}$$

where the trigonometric factors occur at the intersections of the ith and jth rows and columns. Suppressing, for conciseness, the coordinate planes (i,j) in which the rotations take place, G is expressed as:

$$G = G_1^{\mathrm{T}}(\theta_1) G_2^{\mathrm{T}}(\theta_2) \dots G_{p(p-1)/2}^{\mathrm{T}}(\theta_{p(p-1)/2}) \tag{12}$$

The order of the coordinate planes in which the rotations take place is not unique: we mention two possible choices below.

To find expressions for $\partial\rho/\partial\theta_q$ we need only consider the term

$$\begin{aligned} \mathrm{Tr} A^{\mathrm{T}} U B &= \mathrm{Tr} A^{\mathrm{T}} V_0 G D B & (13) \\ &= \mathrm{Tr} A^{\mathrm{T}} V_0 G_1^{\mathrm{T}} G_2^{\mathrm{T}} \dots G_q^{\mathrm{T}}(i_q, k_q, \theta_q) G_{q+1}^{\mathrm{T}} \dots G_{p(p-1)/2}^{\mathrm{T}} D B & (14) \\ &= \mathrm{Tr} H_q G_q^{\mathrm{T}}(i_q, k_q, \theta_q) T_q, & (15) \end{aligned}$$

where

$$H_q = A^{\mathrm{T}} V_0 G_1^{\mathrm{T}} G_2^{\mathrm{T}} \dots G_{q-1}^{\mathrm{T}} \qquad \text{and} \qquad T_q = G_{q+1}^{\mathrm{T}} \dots G_{p(p-1)/2}^{\mathrm{T}} \, DB \tag{16}$$

are the head and tail of the product, and are independent of θ_q. We therefore have

$$\begin{aligned} \frac{\partial}{\partial\theta_q} \mathrm{Tr} H G_q^{\mathrm{T}}(i,k,\theta_q) T = &- \sin\theta_q \sum_{l=1}^{n} H_{li} T_{il} - \cos\theta_q \sum_{l=1}^{n} H_{li} T_{kl} \\ &+ \cos\theta_q \sum_{l=1}^{n} H_{lk} T_{il} - \sin\theta_q \sum_{l=1}^{n} H_{lk} T_{kl}, \end{aligned} \tag{17}$$

where the dependence of H and T on q has been suppressed. It is now straightforward to implement a conjugate gradient scheme to minimize ρ with respect to the $p(p+1)/2$-dimensional vector whose first p elements are the diagonal elements of D and whose remaining elements are the angles θ_q.

Although the conjugate gradient formulation permits minimization of D and G simultaneously, any benefit is by far outweighed by the overhead of repeated evaluations of ρ and its derivatives that are necessitated by the conjugate gradient minimization.

As noted above, there are several ways in which to choose angles θ_q and coordinate plane rotations (i_q, j_q) which represent G. A natural sequence is the order in which the elements of G might be zeroed (with successive premultiplications by G_q) in forming its (trivial) QR decomposition.

That is, as illustrated for a 4×4 matrix in (18), the (i_q, j_q) are chosen to annihilate the lower triangle of G working up each column starting on the left:

$$\begin{bmatrix} \times & \times & \times & \times \\ 3 & \times & \times & \times \\ 2 & 5 & \times & \times \\ 1 & 4 & 6 & \times \end{bmatrix}. \tag{18}$$

Givens elimination of the (i,j)th element requires a rotation in the $(j-1, j)$ plane, which implies rotations in the following sequence of planes:

$$\begin{array}{llllll} (p-1,p), & (p-2,p-1), & (p-3,p-2), & \ldots\ldots & (2,3), & (1,2), \\ (p-1,p), & (p-2,p-1), & (p-3,p-2), & \ldots\ldots & (2,3), & \\ \vdots & & & & & \\ (p-1,p), & (p-2,p-1), & (p-3,p-2), & & & \\ (p-1,p), & (p-2,p-1), & & & & \\ (p-1,p). & & & & & \end{array} \tag{19}$$

It has been our experience, however, that this sequence is numerically inferior to the sequence which makes a single rotation in every pair of coordinate planes:

$$\begin{array}{lllllll} (1,2), & (1,3), & \ldots\ldots & (1,p-2) & (1,p-1), & (1,p) \\ & (2,3), & \ldots\ldots & (2,p-2) & (2,p-1), & (2,p) \\ & & & & & \vdots \\ & & & (p-3,p-2) & (p-3,p-1) & (p-3,p) \\ & & & & (p-2,p-1) & (p-2,p) \\ & & & & & (p-1,p). \end{array} \tag{20}$$

We suspect that the reason for the superiority of (20) over the QR order is that (19) makes repeated rotations in a few planes (e.g. a total of $(p-1)$ rotations in the $(p-1,p)$ plane). A small rotation may be represented by the difference between two or more large clockwise and counterclockwise rotations. In addition to being numerically poorly-conditioned and thus susceptible to round-off error, like the familiar subtraction of two large numbers, it leads to a very flat minimum, close to which a rotation in a particular direction may be balanced by a subsequent rotation in the opposite direction.

2.3 Basis orthogonalization

When B is the identity matrix the classical Procrustes problem seeks the orthogonal matrix, U, which most closely approximates A. If the columns, $\mathbf{a}_i$, of A are regarded as vectors, then the columns of U (that is, the columns of a polar decomposition of A) form a basis for the sub-space spanned by the $\mathbf{a}_i$. Such an orthogonalization of a basis is associated with the names of Löwdin (**?**) and Poincaré. Unlike the Gram-Schmidt procedure, this basis is democratic in that a permutation of the $\mathbf{a}_i$ yields the same basis vectors – though they may be relabelled. If $n > m$ there are more vectors, $\mathbf{a}_i$, than the dimension of the space and the basis is over-determined. In this case an orthonormal basis may be found by solving the Procrustes problem with $B = I_{mn} \in \mathbb{R}^{m \times n}$, the rectangular matrix with ones down the leading diagonal and zeroes elsewhere.

An alternative basis is provided by the V that minimizes the orthogonal Procrustes problem

$$\|A - VDI_{mn}\|^2, \tag{21}$$

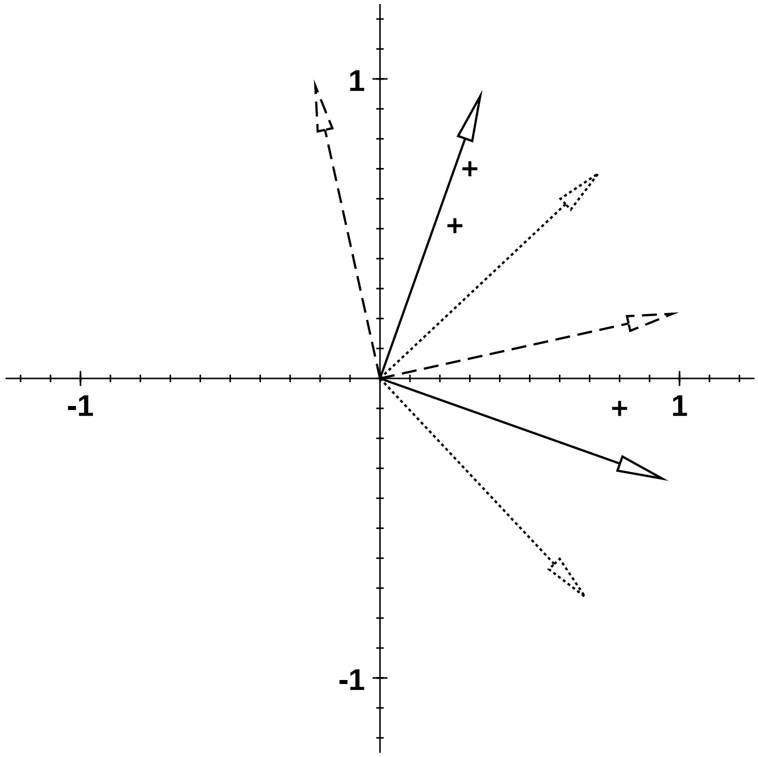

Figure 2: *Two dimensional bases generated from the three points marked with crosses. Basis vectors from Löwdin orthogonalization (dashed), principal components (dotted) and (solid) equation (21).*

with D diagonal. This basis is also invariant under permutations of the a_i, but confers a weighting D_{ii} on the ith basis vector, and the basis tends to be aligned with the original vectors.

Fig. 2 shows 3 two-dimensional vectors (marked with crosses) and the bases found by Löwdin orthogonalization (dashed vectors), equation (21), and the first two left singular vectors of A, which are also known as the principal components (Harman 1960) or empirical eigenfunctions (Sirovich and Everson 1992) of the three vectors. The principal components, $\mathbf{x}_i$, are also the eigenvectors of $A^{\mathrm{T}}A$, the covariance matrix, so the principal eigenvector maximizes the variance of the $\{\mathbf{a}_i\}$ projected onto it: $\sigma_1^2 = \mathbf{x}_1 A^{\mathrm{T}} A \mathbf{x}_1 / \mathbf{x}_1^{\mathrm{T}} \mathbf{x}_1 = (1.07)^2$; the variance of the $\{\mathbf{a}_i\}$ projected onto the second principal component is $\sigma_2^2 = (0.72)^2$, though in this two-dimensional example the second principal component is completely determined by the requirement that it be orthogonal to the first. The diagonal elements of D are 0.72 and 0.065.

3 $A - BU$

In this section we discuss methods for numerically minimizing (6). We point out again that the matrices A, B and U here are different from the matrices with the same names used in the previous section.

3.1 Conjugate gradient tandem

A tandem algorithm analogous to (9), which alternately solves linear equations and the classical Procrustes problem is not possible here. If, as above, U is written in terms of an orthogonal matrix,

V, and a diagonal matrix, D: $U = VD$, (6) becomes

$$\text{minimize } \rho \;=\; \|A - BVD\|^2 \qquad \text{with} \qquad V^{\mathrm{T}}V = I_p. \tag{22}$$

If V is known, then D may again be found by setting the partial derivatives of ρ with respect to the elements of $\mathbf{d} = \mathrm{diag}(D)$ to zero, which yields

$$d_k = \left[A^{\mathrm{T}}BV\right]_{kk} / \left[V^{\mathrm{T}}B^{\mathrm{T}}BV\right]_{kk}. \tag{23}$$

If we regard D as known, (22) is not the classical Procrustes problem, but writing $V = V_0G$ and parameterizing G in terms of $p(p-1)/2$ Givens rotation matrices allows a conjugate gradient minimization. Details of the scheme are messy, but they may be found in MATLAB scripts.[1] Combining these two intermediate steps gives a scheme to minimize (6):

$\mathbf{d}_1 = (1, ..., 1)^{\mathrm{T}}$ (24a)

Obtain (classical) V_1 by minimizing $\|A - BV_1\|^2$ (24b)

Repeat until converged, $i = 2, 3, 4...$ (24c)

  Obtain (eq. (23)) $\mathbf{d}_{i+1}$ by minimizing $\|A - BV_iD_{i+1}\|^2$ (24d)

  Obtain (conjugate gradient) V_{i+1} by minimizing $\|A - BV_{i+1}D_{i+1}\|^2$ (24e)

Go to (24c). (24f)

3.2 Schwarz tandem

In order to avoid using the slow and costly conjugate gradient minimization in (24e), we note that

$$\|A - BVD\|^2 \;=\; \|(AD^{-1} - BV)D\|^2 \;\leq\; \|AD^{-1} - BV\|^2\|D\|^2 \tag{25}$$

One might hope to find an approximation to V_{i+1} as the V that minimizes $\|AD_{i+1}^{-1} - BV\|^2$, namely the orthogonal polar factor of $BD_{i+1}^{-1}A^{\mathrm{T}}$.

The iteration scheme that replaces (24e) with this approximation does not, in general, converge to the minimum of (6), because of the $\|D\|^2$ factor on the right-hand side of (25). However, as Fig. 3 illustrates, the modified iteration, which provides a rapid, but crude approximation, can be followed until $\|A - BV_iD_i\|^2$ no longer decreases, after which the unmodified scheme can be used.

3.3 Newton tandem

Although the conjugate gradient algorithm incorporates derivative information it is relatively slow. In order to speed up (24e), we sacrifice the requirement that each V_i be an orthogonal matrix and linearize the orthogonality condition. This scheme is based on an algorithm due to Watson (Watson 1994) and his presentation is followed here.

We write $U = V(G + H)D$, and regard H as a (small) correction to G. Then, to first order in H the orthogonality requirement $U^{\mathrm{T}}U = D^2$ becomes

$$G^{\mathrm{T}}H + H^{\mathrm{T}}G = I_p - G^{\mathrm{T}}G \tag{26}$$

[1] MATLAB scripts implementing the algorithms presented here are available from http://camelot.mssm.edu/~fme/procrustes

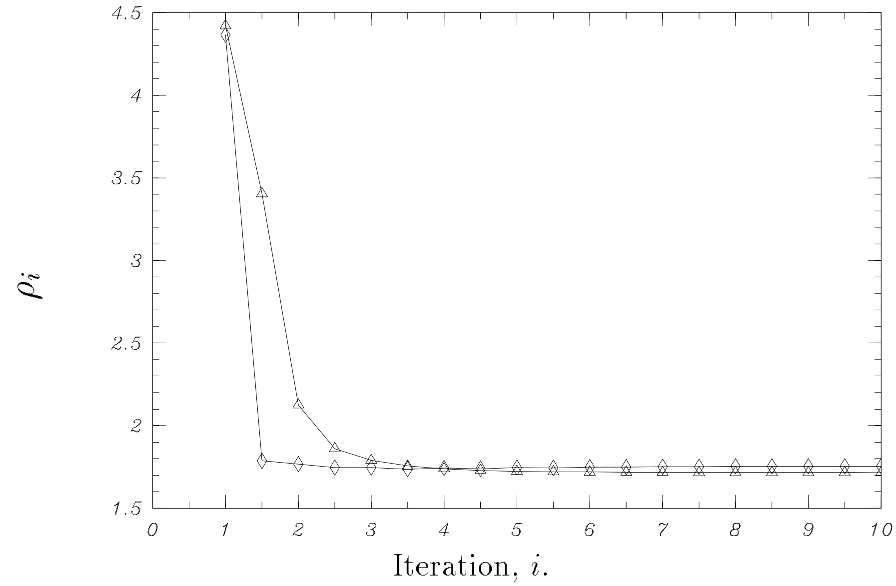

Figure 3: *Convergence of the iteration scheme (24) using (triangles) conjugate gradient minimization for (24e) and (diamonds) a modification using the Schwarz inequality (25). Values of ρ plotted at integer abscissae result from the step (23); those with integer + half abscissae are from the conjugate gradient or Schwarz steps. The modified scheme converges rapidly to a ρ close to the optimum before increasing again. The matrices were $A \in \mathbb{R}^{10\times 3}$ and $B \in \mathbb{R}^{10\times 5}$ with elements drawn from a uniform distribution of random numbers between 0 and 1.*

As Watson observed, H satisfies (26) if and only if

$$H = P + QS, \tag{27}$$

where $Q = G^{-\mathrm{T}}$, $P = (G^{-\mathrm{T}} - G)/2$ and S is any skew-symmetric matrix. The above-diagonal elements of S are now the $p(p-1)/2$ independent variables determining the orthogonal matrix. The formation of Q requires, however, a matrix inversion. To circumvent this let $E = G^{\mathrm{T}}G - I_p$ and take

$$P = -\frac{1}{2}GE \qquad \text{and} \qquad Q = G(I_p - E), \tag{28}$$

which are correct to first order in E.

The independent variables may be expressed as a vector $\mathbf{s} \in \mathbb{R}^{p(p-1)/2}$ formed by concatenating above-diagonal rows of S:

$$\mathbf{s} = (S_{12}, S_{13}, ..., S_{21}, ..., S_{p-1,p})^{\mathrm{T}}, \tag{29}$$

in which case (27) may be written as

$$\mathbf{h} = \mathbf{p} + Z\mathbf{s}, \tag{30}$$

where $\mathbf{h}$ and $\mathbf{p} \in \mathbb{R}^{p^2\times p^2}$ are formed by concatenating the rows of H and P, while $Z \in \mathbb{R}^{p^2\times p(p-1)/2}$ is formed by appropriate rearrangement of the elements of Q.

Neglecting constant terms, the quantity to be minimized becomes

$$2\mathrm{Tr}XHD + \mathrm{Tr}DH^{\mathrm{T}}WHD \tag{31}$$

where $X = (DG^T W - A^T B)$ and $W = V^T B^T BV$. The quadratic term may be expressed as $\mathbf{h}^T M \mathbf{h}$, where W and D have been absorbed into $M \in \mathbb{R}^{p^2 \times p^2}$; the form of M is illustrated for $p = 2$:

$$\begin{bmatrix} W_{11}D_1^2 & 0 & W_{12}D_1^2 & 0 \\ 0 & W_{11}D_2^2 & 0 & W_{12}D_2^2 \\ W_{21}D_1^2 & 0 & W_{22}D_1^2 & 0 \\ 0 & W_{21}D_2^2 & 0 & W_{22}D_2^2 \end{bmatrix}. \tag{32}$$

Let $\mathbf{x}$ be the vector formed by concatenating the columns of X, and $\tilde{D}$ be the diagonal matrix formed from D by repeating its elements p times down the diagonal:

$$\tilde{D} = \text{diag}(D_{11}, ... \quad D_{pp}, D_{11}, \quad ..., D_{pp}). \tag{33}$$

Then (31) becomes

$$2\mathbf{x}^T \tilde{D} Z \mathbf{x} + 2\mathbf{p}^T M Z \mathbf{s} + \mathbf{s}^T Z^T M Z \mathbf{s}. \tag{34}$$

Differentiating with respect to the elements of $\mathbf{s}$ allows $\mathbf{x}$ to be found from a set of linear equations:

$$0 = Z^T \tilde{D} \mathbf{x} + Z^T M \mathbf{p} + Z^T M Z \mathbf{s}. \tag{35}$$

Having found $\mathbf{s}$, G may be updated with H constructed via (27), and the process iterated until convergence.

In fact the updating of $\mathbf{d}$ may usefully be incorporated to give a scheme to solve the full problem (6) as follows:

$\mathbf{d}_1 = (1, ..., 1)^T; \quad G_1 = I_p; \quad V_1 = \text{OPF}(AB^T)$ (36a)

Repeat until converged, $i = 2, 3, 4, ...$ (36b)

  Solve (35) for $\mathbf{s}$ (36c)

  Form H_i from (27) (36d)

  Update $G_{i+1} = G_i + H_i$ (36e)

  (Occasionally) re-orthogonalize G_{i+1} (36f)

    $V_{i+1} = V_i G_{i+1}; \quad G_{i+1} = I_p$ (36g)

    Obtain (eq. (23)) $\mathbf{d}_{i+1}$ by minimizing $\|A - BV_i D_{i+1}\|^2$ (36h)

Go to (36b). (36i)

The scheme, as written, includes an occasional updating of $\mathbf{d}_i$; if the update is omitted on a particular loop of the iteration the estimate for G is polished, eventually becoming orthogonal provided that the scheme converges. Before updating D, however, it is important to ensure that G_{i+1} is orthogonal, so that the orthogonality of the columns of V_i is preserved. Failure to do this results in a final V which does not have orthonormal columns. Reorthonormalization is easily achieved by replacing G_{i+1} by its orthogonal polar factor.

The question of how often to update D depends upon the cost of reorthonormalization. Numerical experiments (see, for example, Fig. 4) suggest that it is more efficient to update D on every traversal of the loop. In this case the construction of (35) is considerably simplified since $Q = I_p$ and $P = 0$, without any approximation. However the convergence rate when updating every traversal is not double that for updates on alternate traversals, so if the convergence is measured per update of D it is more efficient to update less frequently, thus using a better estimate of G_{i+1} in the reorthonormalization.

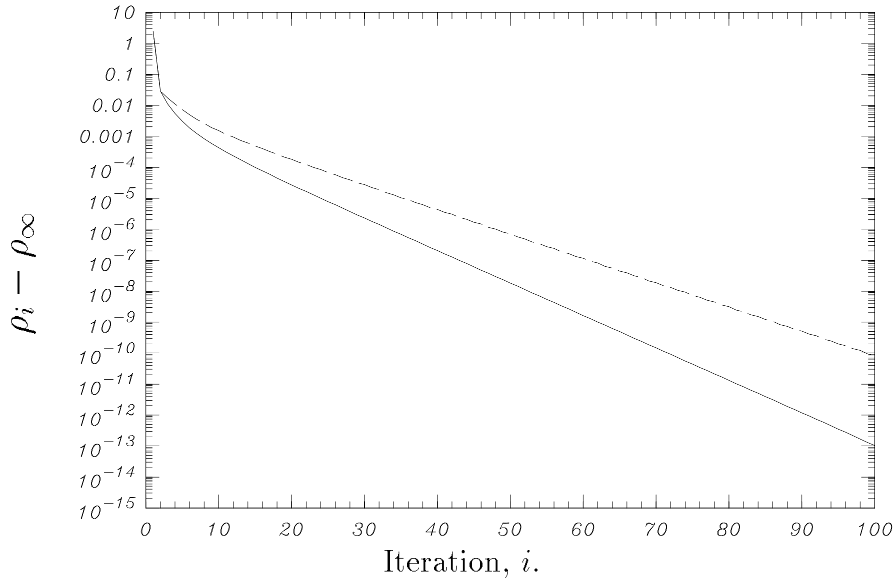

Figure 4: *Convergence of the iteration (36) when D is updated on each traversal of the loop (solid) and on every other traversal (dashed). The differences between ρ at each stage of the iteration and ρ for the fully converged solution are plotted. The matrices were $A \in \mathbb{R}^{20\times5}$ and $B \in \mathbb{R}^{20\times8}$ with elements drawn from a uniform distribution of random numbers between 0 and 1.*

4 Concluding remarks

We have considered the Procrustes problem in which the requirement that the columns of U be orthonormal is relaxed to orthogonality. Various numerical schemes, which converge linearly, have been presented.

Although convergence to a global minimum is not guaranteed, numerical searches on small ($p < 10$) problems have failed to locate better solutions than those achieved by these schemes. This is presumably due, in large measure, to the good initial approximation provided by the orthogonal polar factor when $\mathbf{d} = (1, ..., 1)^{\mathrm{T}}$.

Acknowledgement

I am grateful for discussions with Larry Sirovich and Bruce Knight. This work was supported by the NIMH (NIMH-XXX).

References

Beattie, C. and S. Smith (1992). Optimal matrix approximants in structural identification. *J. Optimization Theory and Applications 74*(1), 23–56.

Everson, R., E. Kaplan, B. Knight, E. O'Brien, D. Orbach, and L. Sirovich (1997). Analysis of optical imaging datasets. *(In preparation)*.

Golub, G. and C. Van Loan (1983). *Matrix Computations*. Oxford: North Oxford Academic.

Gower, J. (1984). Multivariate analysis: Ordination, multidimensional scaling and allied topics. In E. Lloyd (Ed.), *Statistics*, Volume VI of *Handbook of applicable mathematics*, pp. 727–781. Chichester: John Wiley.

Green, B. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika 17*(4), 429–440.

Harman, H. (1960). *Modern Factor Analysis*. Chicago: Univ. Chicago Press.

Higham, N. (1986). Computing the polar decompostion – with applications. *SIAM J. Sci. Stat. Computing 7*(4), 1160–1174.

Horn, R. and C. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

Sirovich, L. and R. Everson (1992). Analysis and management of large scientific databases. *International Journal of Supercomputing Applications 6*(1), 50–68.

Watson, G. (1994). The solution of orthogonal procrustes problems for a family of orthogonally invariant norms. *Advances in Computational Mathematics 2*, 393–405.