*Multivariate Behavioral Research*, 26 (1), 179-197


# Theory & Practice of Analytic Hyperplane Optimization

William W. Rozeboom
University of Alberta

There are many procedural alternatives by which subjective rotation to oblique simple structure can be emulated analytically, some rather more effective than others. This article develops a theoretical framework for comprehending these, and reports performance tests for some major variants thereof when implemented within the HYBALL rotation program described elsewhere.

The HYBALL program for factor rotation to oblique simple structure, described in Rozeboom (1991), has been designed primarily to provide invariance under rotation of selected factor axes/subspaces. But it also advances the art of analytic emulation of subjective rotation. Here's your chance to learn more about this exciting development.

For any to-be-rotated factor pattern $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m]$ of variables $Y = (y_1, ..., y_n)$ on factors $F = (f_1, ..., f_m)$ (i.e., the $i$th element of column vector $\mathbf{a}_j$ is the initial loading of item $y_i$ on factor $f_j$), consider the pattern shift that results from replacing $f_1$ by $\tilde{f}_1 = f_1 + wf_2$ for an arbitrary rotation coefficient $w$. As detailed in Rozeboom, 1991, this is a rotation just of $f_1$ in the $f_1/f_2$ plane that leaves unaltered all pattern coefficients except the column $\mathbf{a}_2$ of loadings on $f_2$. Specifically, prior to renormalizing the rotated factor's variance (which merely multiplies pattern column $\mathbf{a}_1$ by a scaling constant), the only pattern change that results from this single-plane shift in $f_1$ is replacement of the pattern column $\mathbf{a}_2$ on $f_2$ by

$$\tilde{\mathbf{a}}_2 = \mathbf{a}_2 - w\mathbf{a}_1 \quad (\tilde{f}_1 = f_1 + wf_2).$$

(See Rozeboom, 1991, Equation 8 and Figure 1. Take care to appreciate how this shift of $f_1$ in the $f_1/f_2$ plane modifies the distribution of loadings not on $f_1$ but on $f_2$. Justification for omitting renormalization of $\tilde{f}_1$'s variance here is sketched in the Appendix.) Given that we wish to find the strongest hyperplanes, our task is to solve for the $w$ that yields an optimal array of near-zero elements in $\tilde{\mathbf{a}}_2$. In applications, of course, such single-axis repositionings are to be aggregated for all moveable factors over all planes, and iterated through as many rotation cycles as may be needed for convergence to global optimum. But the mechanics of global iteration (see Rozeboom, 1991, Appendix) are not of present concern.

Here, our focus is how to locate the best hyperplane for $f_2$ in the pattern just on factor pair $(f_1, f_2)$.

## *Theory*

Analytic identification of the best $w$ for axis-shift $\tilde{f}_1 = f_1 + wf_2$ evidently requires above all some choice of a criterion measure $\gamma$ on pairs of pattern columns to appraise the ideality of rotated planar pattern $(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$. It then remains to contrive an efficient computation of the $w$ that optimizes $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$, that is, $\gamma(\mathbf{a}_1, \mathbf{a}_2 - w\mathbf{a}_1)$; but first concern must be to articulate a class of reasonable alternatives for $\gamma$. Because our target is simple structure, we want $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ to become increasingly optimal as each element $\tilde{a}_{i2}$ of $\tilde{\mathbf{a}}_2$ approaches zero; so $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ is to be a monotone function of each $|\tilde{a}_{i2}|$ $(i = 1, ..., n)$. But what function shape beyond that? And is its value to be determined *just* by the array $\{|\tilde{a}_{i2}|\}$ of rotated pattern magnitudes on axis $f_2$ (so that $\mathbf{a}_1$ occurs vacuously in $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ ), or are subtler aspects of pattern configuration $(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ also to be relevant?

To close in on details for $\gamma$, we stipulate first that this is to be a real-valued function optimal at an extremum. But should the latter be a maximum or a minimum? If we think of each $|\tilde{a}_{i2}|$ as an "error" in the $i$th pattern point's fit to the provisional $f_2$-hyperplane, and take $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ to measure how distasteful is this array of misfits on the whole, $\gamma$ will be optimally at minimum (most naturally zero) when each $|\tilde{a}_{i2}|$ is zero. Yet as embodied in previous approaches to analytic hyperplane optimization, notably Maxplane (Cattell & Muerle, 1960; Eber, 1966) and Functionplane (Katz & Rohlf, 1974, 1975), we can alternatively let $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ appraise the concentration of points in this plane around the $f_2$-hyperplane. In this latter case, $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ reaches its optimal maximum (most naturally $n$) when all rotated coefficients $(\tilde{a}_{i2}\}$ on $f_2$ are zero, and approaches zero as the hyperplane's neighborhood becomes empty. However, any such concentration-near-hyperplane measure can be translated as poorness-of-fit simply by subtracting it from its maximum, whereas the converse conversion does not work for loss-functions that are unbounded. So hyperplane misfit is the broader conception.

Consider, therefore, the class of misfit measures on $(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ that are minimal at zero when $\tilde{\mathbf{a}}_2 = \mathbf{0}$ and are increasing in each $|\tilde{a}_{i2}|$. Simplest of these is just

$$\sum_{i=1}^{n} |\tilde{a}_{i2}|$$

— which, however, is useful mainly to exhibit inadequacies. For one, its explicit dependence on the absolute-value operator is analytically infelicitous. More importantly, although we find it most algebraically appropriate for $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$ to be a sum of terms $\gamma_i(a_{i1}, \tilde{a}_{i2})$ contributed by the individual pattern points, there

is no reason why each $\gamma_i(a_{i1}, \tilde{a}_{i2})$ needs be linearly proportional to $|\tilde{a}_{i2}|$; instead, its increase with increasing $|\tilde{a}_{i2}|$ can just as well be either positively accelerated (as in traditional least-squares curve fitting) or decelerated with or without an upper bound. Finally,

$$\sum_{i=1}^{n} |\tilde{a}_{i2}|$$

makes no provision for each pattern point's contribution to the total-misfit measure being distinctively weighted to reflect pattern desiderata beyond this point's bare distance from the $f_2$-hyperplane. Thus, we may wish to assign zero weight to points that are evidently salient on $f_2$; and we might also consider small loadings on $f_2$ to be worth less in points near the $f_1/f_2$ plane's origin than in points that are strongly salient on $f_1$.

Accordingly, let us take as our generic class of misfit measures — not comprising all mathematical possibilities but surely broad enough to include most that are analytically tractable — the ones expressible as a function of hyperplane repositioning in form

$$E_\psi(w) = \sum_{i=1}^{n} c_i \times \psi[\varepsilon_i(w)], \tag{1}$$

[1]

wherein each $\varepsilon_i(w)$ is a nonnegative error term generated from the rotated pattern point $(a_{i1}, \tilde{a}_{i2})$ by choice of rotation coefficient $w$, $\psi(\varepsilon)$ is some fixed growth-function of $\varepsilon$, and each $c_i$ is some fixed nonnegative weight. We shall further take each $\varepsilon_i(w)$ to have the more determinate specification

$$\varepsilon_i(w) = (\tilde{a}_{i2}/b)^2 = [(a_{i2} - wa_{i1})/b]^2 \tag{2}$$

where $b$ is a hyperplane-bandwidth parameter, say between .10 and .20, such that point $i$ is construed to be in the rotated $f_2$-hyperplane just in case $|\tilde{a}_{i2}| \leq b$. But details of $\varepsilon_i(w)$ do not yet matter beyond the important observation that when $\psi$ is the Identity function, solving [1, 2] for the $w$ that minimizes $E_\psi(w)$ is simply a weighted least-squares regression solution for $w$ in $w\mathbf{a}_1 \approx \mathbf{a}_2$. That is, in special case $\psi(\varepsilon) = \varepsilon$, (1, 2) simplifies to

$$E_*(w) = \sum_{i=1}^{n} c_i\varepsilon_i(w) = \sum_{i=1}^{n} b_i^2\tilde{a}_{i2}^2 \qquad (b_i =_{def} \sqrt{c_i}/\,b), \tag{3}$$

$$= \sum_{i=1}^{n} ([b_i a_{i2}] - w[b_i a_{i1}])^2$$

[1] $E_\psi(w)$ is essentially a specification of $\gamma(\mathbf{a}_1, \mathbf{a}_2\text{-}w\mathbf{a}_1)$, except that the pattern coefficients which are explicit arguments in the latter are only implicit in the former — primarily through definition Equation 2 of $\varepsilon_i$, but also prospectively as constituents of weights $(c_i)$ (cf. later Equation 12).

which is the sum-of-squares loss resulting from first replacing unrotated pattern $(\mathbf{a}_1, \mathbf{a}_2)$ by its weighting $(\hat{a}_{i1},\hat{a}_{i2}) =_{def} (b_i a_{i1}, b_i a_{i2})$ $(i = 1, 2; ..., n)$ and then approximating $\hat{\mathbf{a}}_2$ by $w\hat{\mathbf{a}}_1$. The $w$ that optimizes fit in this case is of course the classic least-squares solution

$$(4) \qquad w_{opt} = (\hat{\mathbf{a}}_1'\hat{\mathbf{a}}_1)^{-1}\hat{\mathbf{a}}_1'\hat{\mathbf{a}}_2 = (\sum_{i=1}^{n} b_i a_{i1} a_{i2})/(\sum_{i=1}^{n} b_i^2 a_{i1}^2)$$

$$= (\sum_{i=1}^{n} c_i a_{i1} a_{i2})/(\sum_{i=1}^{n} c_i a_{i1}^2).$$

In fact, (3, 4) is HYBALL's basic *step-down regression* solution for provisional hyperplanes, with binary weights $(c_i)$ set according to whether point $i$ lies in the search window at the moment of application. In this procedure, $c_i$ is 1 (otherwise 0) just in case the angle of point $i$ to the to-be-rotated axis $f_1$ is less than an adjustable value $\beta$. Shifts of $f_1$ using Equation 4 are iterated with $\beta$ starting large, say $\beta = 60°$, but decreased after each iteration until $\beta$ reaches a lower limit of 15° or thereabouts. Despite its lack of finesse, basic step-down regression often works nearly as well as more sophisticated solutions.

Another instance of hyperplane fitting by a criterion having the form of Equation 3 is Direct Quartimin (Jennrich & Sampson, 1966), wherein $c_i$ is proportional to $a_{i1}^2$. More will be said about this case in the Appendix.

A large objection to Equation 4, however, is its zeal in trying to minimize the deviancy of outliers, due to $\varepsilon_i(w)$'s positive acceleration in $|\tilde{a}_{i2}|$, even though outliers are the most conspicuous salients on $f_2$ and should hence be given little say in positioning the $f_2$-hyperplane. Assigning zero weights to evident outliers is a workable solution to this problem, but rather a crude one. Far more elegant is the graded analytic disregard for outliers that suitable $\psi$ can accomplish in Equation 1 by making $\psi[\varepsilon_i(w)]$ a *decelerated* (i.e., convex) increasing function of $|\tilde{a}_{i2}|$, preferably one having an upper bound. For then, when $|\tilde{a}_{i2}| > |\tilde{a}_{j2}|$, even though $\psi[\varepsilon_i(w)]$ still adds more to $E_\psi(w)$ in Equation 1 than does $\psi[\varepsilon_j(w)]$, the *change* in $\psi[\varepsilon_i(w)]$ produced by a small change in $w$ is now less than the change in $\psi[\varepsilon_j(w)]$; so for small adjustments of the provisional $f_2$-hyperplane selected by an approximation $w$ to $w_{opt}$, points whose misfits $|\tilde{a}_{i2}|$ by the provisional hyperplane are smallest will have the strongest voice in appraising the merit of this shift while outliers for which $\psi[\varepsilon_i(w)]$ is near $\psi$'s limit will be virtually mute. Under such a $\psi$, differential item weights in Equation 1 may still have some purpose; but they are no longer needed to exclude outliers.

Instead of requiring $\psi(\varepsilon_i)$ to be convex throughout, we can achieve the same outlier suppression by giving this a sigmoid shape under which deceleration of its increase in $|\tilde{a}_{i2}|$ does not take over until $|\tilde{a}_{i2}|$ leaves the hyperplane band. The most extreme example is Maxplane's hyperplane counting which, rewritten as

a misfit measure, takes $\psi(\varepsilon)$ to be 0 if $\varepsilon \leq 1$ and 1 if $\varepsilon > 1$. Functionplane, too, has generally sigmoid misfit.

Three practical questions remain: (a) What are good functions to choose for $\psi$ in Equation 1? (b) Have we any rationale for differential weights $\{c_i\}$ in Equation 1 even when $\psi$ is bounded? And (c) how can we best solve Equations 1/2 for the $w_{opt}$ that minimizes $E_\psi(w)$ given our choice of $\psi$? I shall speak to (c) first, for I have two methods to recommend — brute-force scanning, and polished step-down regression — and both are largely indifferent to the specifics of $\psi$.

Brute-force scanning of Equations 1/2 for $w_{opt}$ is simply computation of $E_\psi(w)$ over a sufficiently dense sampling of $w$-values and returning the $w$ for which $E_\psi(w)$ is smallest. If we limit arctan($w$) to the range, say, ±75°, 26 computations of $E_\psi(w)$ for equally spaced arctan($w$) should identify arctan($w_{opt}$) within 6°, after which another 25 or so finer-spaced computations of $E_\psi(w)$ in this vicinity finds $w_{opt}$ to 2nd-decimal precision. Programming details are entirely routine, mainly matters of how $\psi$ is to be specified by what control parameters in what determinate function-form, and how, once specified, it can be computed most efficiently. The main practical complication is that if computation of $\psi$ is sluggish, its approximately $50 \times n$ applications in each solution for $w_{opt}$, multiplied by the number of coefficients to be found for all shifts in all factor planes times the number of iteration cycles, can be rather time-consuming.

But when $\psi$ has a continuous 1st derivative, there is also an alternative computation that can generally find $w_{opt}$ in *much* less time than by brute-force scanning. This is to approximate Equations 1/2 by an instance of Equation 3 wherein each item weight is the 1st derivative of $c_i \times \psi$ evaluated at the item's current value of $\varepsilon_i(w)$. Specifically, the 1st derivative of $E_\psi(w)$ in Equation 1 with respect to $w$ is

$$E_\psi'(w) = \sum_{i=1}^{n} c_i \times \psi'[\varepsilon_i(w)] \times \varepsilon_i'(w), \tag{5}$$

while the 1st derivative of Equation 3 is

$$E_*'(w) = \sum_{i=1}^{n} c_i \times \varepsilon_i'(w). \tag{6}$$

If we replace each fixed weight $c_i$ in Equation 3 by $c_i$ times $\psi'[\varepsilon_i(w_o)]$ at the current estimate $w_o$ of $w_{opt}$ and write $E_o(w)$ for this special weighting of $E_*(w)$, the slope of $E_o(w)$ at $w = w_o$ is thus the same as that of $E_\psi(w)$. So $E_\psi(w)$ must have a local minimum — which moreover should be its global minimum if $w_o$ is fairly close to $w_{opt}$ — in the same direction from $w_o$ as the minimum of $E_o(w)$ and quite possibly in its vicinity. This suggests that starting from an initial

approximation $w_o$ to $w_{opt}$, the value $w_1$ of $w$ that minimizes $E_o$, namely, the value computed by Equation 4 with item weights $\{c_i \times \psi'[\varepsilon_i(w_o)]\}$, should be an efficiently improved approximation to $w_{opt}$. There is some risk of overshooting; but to minimize that we can let our revised approximation to $w_{opt}$ be $pw_1 + (1 - p)w_o$ for some damping parameter $p$ in the unit interval. (In fact, overshooting virtually never occurs for functions in the class implemented by HYBALL, so by rights $p \geq 1$ should be most efficient. But for reasons not clear to me, $p \approx .5$ generally yields somewhat higher hyperplane counts when iterated to global solutions in difficult problems than does $p$ near 1.) I shall call this method *polished* step-down regression in contrast to its basic (unpolished) version that delivers the point of departure for polishing. More technically, HYBALL programs step-down regression with a parametric limit *lim* on the number of local polish iterations (*strokes*) allowed per factor pair on any one global iteration cycle. (*Lim* = 10 is a good default setting; most local polish iterations converge within that when $p \approx .5$ even in large problems.) Basic step-down regression is then special case *lim* = 0.

What should we choose for $\psi$? The function-form proposed by Katz & Rohlf (1974), namely,

$$(7) \qquad \psi(\varepsilon) = 1 - 1/\mathrm{Exp}(\varepsilon^r), \qquad \psi'(\varepsilon) = r[1 - \psi(\varepsilon)]\varepsilon^{r-1} \quad (r > 0),$$

with $r$ a control parameter having default value $r = 1$, is very attractive mathematically. But Equation 7 is not very parametrically flexible; and worse, Exponentiation is slow to compute. So consider instead the class of growth functions whose 1st derivatives are delimited by suitable constraints on parameters $a, b, c, r, s$ in

$$\psi'(\varepsilon) = \begin{cases} b - a\varepsilon^r & \text{if } \varepsilon \leq 1 \\ c/\varepsilon^s & \text{if } \varepsilon > 1 \end{cases}.$$

We require $\psi'(\varepsilon) \geq 0$ for all $\varepsilon$; so $b \geq 0$, and also $r \geq 0$ unless $a \leq 0$.[2] And $\psi'(1) = 1$ is a convenient scaling stipulation, whence $b = 1 + a$ and $a \geq -1$, as well as $c = 1$ unless we allow $\psi'(\varepsilon)$ to be discontinuous at $\varepsilon = 1$ (a permissible generality that we shall forego). This gives us the class of functions whose 1st derivatives are

[2] $b \geq 0$ is immediate from $\psi'(0) \geq 0$. And were $r$ to be negative when $a$ is positive, $\psi'(\varepsilon)$ would become negative for sufficiently small $\varepsilon$. All other combinations of $a$ and $r$ are admissible by this criterion; however, $r < -1$ will soon prove inadmissible on other grounds.

$$(8) \qquad \psi'(\varepsilon) = \begin{cases} 1 + a(1 - \varepsilon^r) & \text{if } \varepsilon \le 1 \\ 1/\varepsilon^s & \text{if } \varepsilon > 1 \end{cases} \qquad (a \ge -1;\ r \ge 0 \text{ unless } a \le 0);$$

The functions that have Equation 8 derivatives ($r \neq -1, s \neq 1$) and for which also $\psi(0) = 0$ (which cannot be satisfied if $r < -1$) are

$$(9) \qquad \psi(\varepsilon) = \begin{cases} (1 + a)\varepsilon - a(r + 1)^{-1}\varepsilon^{r+1} & \text{if } \varepsilon \le 1 \\ d - 1/(s - 1)\varepsilon^{s-1} & \text{if } \varepsilon > 1 \end{cases} \qquad (r > -1),$$

wherein the value of $d$ required for $\psi(\varepsilon)$ to be continuous at $\varepsilon = 1$ is

$$d = s/(s - 1) + ar/(r + 1).$$

For $s = 1$, the integral of Equation 8's $\varepsilon > 1$ leg is $\psi(\varepsilon) = d + ln(\varepsilon)$; so Equation 9 increases in $\varepsilon$ without limit just in case $s \le 1$. That is, we need $s > 1$ for $\psi$ to be bounded.

For small integers $r$ and $s$, functions having the form of Equations 8/9 are computationally quite fast and provide considerable diversity of growth curves, either convex in $\varepsilon$ throughout when $ar > 0$, or sigmoid when $a < 0$ with inflection point $\varepsilon = 1$. Equation 9 resembles Equation 7 most closely when $a = -1$. However, with $r \ge 0$ (see *Note 2*, below, on special setting $r < 0$) this makes $\psi'(\varepsilon)$ approach 0 as $\varepsilon$ approaches 0, which seems unattractive for polished step-down regression. If intuition is correct to insist that points near center in the hyperplane band should carry more weight in hyperplane positioning than points at its edge, we should prefer $a > 0$. And in that case we can also stipulate $s = ar$ to contrive (admittedly for no compelling reason beyond mathematical elegance) that Equation 9's 2nd derivative, too, is continuous throughout. On the other hand, if we don't much care how close points are to hyperplane center so long as they are in the hyperplane band somewhere, the way to go is $a = -1$ with $r$ at least 1 and preferably greater.

There is, of course, rather more to Equation 9's parametric flexibility than just its curvature within the hyperplane band. Another cogency is that for $\varepsilon > 1$, misfit approaches its limit with increasing rapidity as $s$ becomes large. But even more prima facie important is how $\psi$'s range from $\psi(0)$ to $\psi(\infty)$ divides between the rise $h_{in}$ from hyperplane center to edge and its remaining rise $h_{out}$ beyond that. To make this comparison, observe from Equation 9 that

$$\begin{aligned} h_{in} &= \psi(1) - \psi(0) = 1 + a - a/(r + 1) = 1 + ar/(r + 1), \\ h_{out} &= \lim_{\varepsilon \to \infty} \psi(\varepsilon) - \psi(1) = 1/(s - 1); \end{aligned}$$

so $h_{in}$ in proportion to $h_{out}$ is

$$h_{in}/h_{out} = (s - 1)[1 + ar / (r + 1)] \approx \begin{cases} (s - 1)/(r + 1) & \text{if } a \approx -1 \\ (s - 1)(1 + a) & \text{if } r \geq 2 \end{cases} .$$

This does not tell us what $h_{in}/h_{out}$ values to prefer; but it does facilitate inquiry whether differences in the quality of results under varied parameter settings can be attributed substantially to variation in this ratio.

*Note 1*

For very large settings of parameter $b$ in Equation 2 specification of $\varepsilon$, say $b > 1$, $b$ will exceed almost every $\tilde{a}_{i2}$ encountered in practice so that only the first leg of Equation 9 is operative. Hence Equation 9 in effect includes measures $\psi(\varepsilon)$ that are positively accelerated throughout, even though use of these for hyperplane search would ordinarily be unmotivated. More generally, large $b$ allows the first leg of Equation 9 to be operative throughout regardless of its curvature.

*Note 2*

The first leg of Equation 9 under $a = -1$ and large hyperplane-bandwidth $b$ (see Note 1) holds special theoretical interest for fractional $r$ between 0 and -1. Reparameterizing this range by $\nu = (1 + r)^{-1}$ gives us

$$\psi(\varepsilon) = \nu\varepsilon^{1/\nu} \quad (b > 1) \tag{10}$$

as a special subclass of Equation 9 wherein $\psi(\varepsilon)$ is proportional to the $\nu$th root of $\varepsilon$ for any positive $\nu$. And Equation 10 in turn specifies Equations 1/2 for the $f_2$-hyperplane misfit resulting from axis shift $\tilde{f}_1 = f_1 + wf_2$ as

$$E_{**}(w) = \sum_{i=1}^{n} b_i \times |\tilde{a}_{i2}|^{2/\nu} \qquad (b_i = c_i / b^{2/\nu},\ \nu > 0) . \tag{11}$$

$$= \sum_{i=1}^{n} b_i \times |a_{i2} - wa_{i1}|^{2/\nu}$$

$E_{**}$ is a generalization of Equation 3 whose increase in $|\tilde{a}_{i2}|$ is positively accelerated if $\nu < 2$, linear if $\nu = 2$, and negatively accelerated if $\nu > 2$. What is most theoretically appealing about Equation 11 is that hyperplane bandwidth $b$ affects only the scaling of $E_{**}$ with no impact whatever on its solution for $w_{opt}$,

while the hyperplane identified by $w_{opt}$ is invariant under rescalings of $\mathbf{a}_1$ and $\mathbf{a}_2$. Specifically, for any fixed item weights ($c_i$) and any rescaling $(\mathbf{a}_1^*, \mathbf{a}_2^*) = (s_1\mathbf{a}_1, s_2\mathbf{a}_2)$ of pattern configuration $(\mathbf{a}_1,\mathbf{a}_2)$ (positive $s_1$ and $s_2$), the value $w_{opt}^*$ of $w$ that minimizes $E_{**}(\mathbf{a}_1^*, \mathbf{a}_2^*, w)$ is $w_{opt}^* = (s_1/s_2)w_{opt}$, so that $\tilde{a}_{i2}$ $(= a_{i2} - w_{opt}a_{i1})$ is zero just in case $\tilde{a}_{i2}^* = a_{i2}^* - w_{opt}^* a_{i1}^* = s_2a_{i2} - (s_2/s_1) \times w_{opt} \times (s_1a_{i1}) = s_2\tilde{a}_{i2}$ is zero. In short, rotations based on misfit functions of Equation 11 form have the extremely desirable property of being *scale invariant*, or more precisely factor-scale invariant, with concommitant indifference to hyperplane bandwidth. And although the price of this is unboundedness of Equation 11 in $|\tilde{a}_{i2}|$, the influence on axis positioning of points far from hyperplane center can be emasculated as thoroughly as we please by taking $v$ sufficiently large.

*Note 3*

We can take

$$\psi(\varepsilon) = v\varepsilon^{1/v} \quad \text{if } \varepsilon \leq 1$$

as an additional alternative for the first leg of Equation 9 even when hyperplane-width parameter $b$ is small. I have added this extension to HYBALL, with the expectation but not requirement that it be used with large $b$, by letting negative settings of integer control parameter $r$ call its variant $v = 1 + |r|$. However, I have found no applications in which this option improves on $r \geq 0$; and it increases the computation time of polish strokes and brute-force scanning by a factor of three.

*Note 4*

If called for adjustment during a HYBALL run, the parameters here written $a$, $r$, $s$, and $p$ are identified on screen as CV, JA, JB, and PD, respectively.

*Practice*

I have tested parametric variations in these hyperplane-search methods by HYBALL rotation of eight empirical problems ranging from 11 to 86 variables and from 3 to 10 factors. The variants compared were (a) unpolished step-down regression (STEP) versus polished step-down regression with *lim* = 10 (POL) versus brute-force scanning (SCAN); (b) hyperplane-bandwidth target $b = .15$ versus $b = .10$; (c) Varimax start versus no Varimax start; (d) misfit curvature within the targeted hyperplane band at settings $a = -1, 0, 1, 2, 3$; (e) power levels $r = s = 2$ versus $r = s = 4$; and (f) Equation 9 versus Equation 7 for the form of

$\psi$. In the majority of these problems, procedural variations affected the resultant hyperplane counts within the ±.15 band by at most a percentage point or two. (Differences within the ±.10 band were somewhat more pronounced.) But even these small differences were near-perfectly consistent with the effects summarized in Table 1 for the three problems that were most responsive to method detail. These are (a) the Holtzinger pattern (Harman, 1960) of 24 variables on four factors, (b) the pattern on 10 factors extracted from the correlations published by Thurstone (1938) for 57 PMA tests, and (c) the pattern on six factors extracted from data on 86 thinking-style scales in an unpublished study by my colleagues W. Baker & L. Mos.

Before commenting on the Table 1 comparisons, let me first note findings that don't need tabling, namely, computational efficiency, the virtues of Equation 9 versus Equation 7, benefits of Varimax start, and some uncertainties about the most effective use of POL. Regarding Varimax, only one of the problems studied showed any gain from commencing with Varimax prerotation, and even that mattered only for STEP. (This was the Holtzinger pattern, for which Varimax alleviates a hyperplane ambiguity that causes persistent instability in one plane for STEP and POL.) Even so, Varimax start takes negligible time, seldom does any harm, and sometimes leads to a rotated pattern interestingly different from the one reached without it. So it remains a default option in HYBALL.

Efficiency-wise, Equation 7 took over three times as long to compute as did Equations 8/9, both in brute-force scanning and for POL weighting. Moreover, Equation 7 never achieved better hyperplane counts than did its nearest Equation 9 counterpart but tended (unreliably) to do slightly worse. So you'll hear no more about Equation 7 here. Secondly, polishing step-down regression by Equation 8 under *lim* = 10 increased STEP time by only about 25%, whereas brute-force scanning under the corresponding Equation 9 took roughly *fifteen times* that long. Subsequent to running these comparisons I have managed to improve SCAN's efficiency, but it still exceeds POL time by almost an order of magnitude.

However, POL's advantage is scarcely unequivocal. For SCAN prevailingly achieves higher hyperplane counts in complex problems than does POL; and although the difference needn't be appreciable, occasional vagaries in POL's behavior may vitiate much of its per-cycle speed superiority. Convergence in global iteration under SCAN is generally sharper and more rapid than under POL, nor does the hyperplane count under POL always increase monotonically with the number of iterations. Moreover, although POL damping $p = .5$ has consistently proved superior to $p$ much larger or smaller than this, the chance that this will not always be true seems high enough to encourage users to explore other $p$-settings as well. Such complications do not arise under SCAN.

Table 1

Percent of loadings in each of the three factor patterns rotated by HYBALL into the ±.15 and ±.10 hyperplane bands, variously using unpolished step-down regression (STEP), polished step-down regression with *lim* = 10 (POL), and brute-force scanning (SCAN) under various settings of misfit parameter b (hyperplane-bandwidth target), *a* (misfit curvature within the target hyperplane-band), and powers *r* and *s*.

| | | | | *A* (Holtzinger), 24 × 4 | | | | | | *B* (Thurstone), 57 × 10 | | | | | | *C* (Baker/Mos), 86 × 6 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | % rotated loadings in band | | | | | | % rotated loadings in band | | | | | | % rotated loadings in band | | | | | |
| | | | | ±.15 | | ±.10 | | averaged | | ±.15 | | ±.10 | | averaged | | ±.15 | | ±.10 | | averaged | |
| | | | | under each target *b*-setting | | | | | | under each target *b*-setting | | | | | | under each target *b*-setting | | | | | |
| Method | *a* | *r* | *s* | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 | .15 | .10 |
| STEP | - | - | - | 57.3 | 57.3 | 44.8 | 44.8 | 51.0 | 51.0 | 60.5 | 58.1 | 48.2 | 50.2 | 54.4 | 54.2 | 67.6 | 69.2 | 57.6 | 56.8 | 62.6 | 63.0 |
| POL | -1 | 2 | 2 | 57.3 | 57.3 | 41.7 | 43.8 | 49.5 | 50.6 | 68.2 | 61.9 | 42.6 | 56.4 | 55.4 | 59.2 | 73.8 | 73.1 | 53.9 | 61.0 | 63.8 | 67.0 |
| SCAN | -1 | 2 | 2 | 57.3 | 56.2 | 37.3 | 49.0 | 47.3 | 52.6 | 70.9 | 61.8 | 43.0 | 58.5 | 57.0 | 60.3 | 76.0 | 73.4 | 56.2 | 63.2 | 66.1 | 68.3 |
| POL | +1 | 2 | 2 | 56.2 | 55.2 | 44.8 | 44.8 | 50.5 | 50.5 | 63.7 | 59.6 | 52.3 | 52.3 | 58.0 | 56.0 | 70.9 | 70.5 | 58.7 | 58.5 | 64.8 | 64.5 |
| SCAN | +1 | 2 | 2 | 55.2 | 45.8 | 46.9 | 42.7 | 51.0 | 44.2 | 63.7 | 60.5 | 55.8 | 53.7 | 59.8 | 57.1 | 73.3 | 71.7 | 61.4 | 61.6 | 67.4 | 66.7 |
| POL | 0 | 2 | 2 | 55.2 | 54.2 | 43.8 | 46.9 | 49.5 | 50.6 | 64.4 | 61.1 | 48.9 | 53.2 | 56.6 | 57.2 | 71.9 | 70.7 | 58.3 | 58.9 | 65.1 | 64.8 |
| POL | +2 | 2 | 2 | 57.4 | 55.3 | 45.7 | 46.8 | 51.6 | 51.0 | 62.8 | 60.2 | 52.8 | 51.4 | 57.8 | 55.8 | 70.9 | 70.3 | 59.3 | 57.8 | 65.1 | 64.0 |
| POL | +3 | 2 | 2 | 55.3 | 55.3 | 47.9 | 46.8 | 51.6 | 51.0 | 63.0 | 60.0 | 53.7 | 51.1 | 58.4 | 55.6 | 70.7 | 69.8 | 59.3 | 58.1 | 65.0 | 64.0 |
| POL | +1 | 4 | 4 | 56.2 | 54.2 | 46.9 | 45.8 | 51.6 | 50.0 | 62.5 | 59.5 | 52.6 | 52.5 | 57.6 | 56.0 | 71.3 | 69.6 | 59.1 | 58.1 | 65.2 | 63.8 |
| POL | -1 | 4 | 4 | 58.3 | 57.3 | 42.7 | 46.9 | 50.5 | 52.1 | 67.7 | 60.7 | 44.0 | 56.3 | 55.8 | 58.5 | 75.0 | 70.7 | 55.0 | 60.9 | 65.0 | 65.8 |

And now to Table 1's hyperplane ratings. These exhibit the rather interesting effect of within-hyperplane misfit curvature $a$, compares STEP/POL/SCAN, and looks briefly at variation in powerings $r$ and $s$. The conclusions:

1. Although POL clearly improves upon STEP when that is possible, and with well-chosen $p$ decently approaches the hyperplane counts achieved by SCAN under the same parameter settings,[3] SCAN wrings the highest hyperplane counts out of complex patterns. Moreover — an important point not shown in Table 1 but already noted — global iteration reaches convergence more rapidly and reliably under SCAN than under POL.

2. At each hyperplane-bandwidth target $b = .15$ and $b = .10$, curvature setting $a = -1$ consistently yields a higher count within the target band than does $a = +1$, with $a = 0$ intermediate but generally closer to $a = +1$. But under $b = .15$, $a = +1$ obtains higher counts for inner band $\pm.10$ than does $a = -1$, with $a = 0$ again intermediate but closer to $a = +1$. This is as might be expected; for $a = +1$ works harder than does $a = -1$ to center points within the target band, and should accordingly be somewhat less attentive to herding points at band edge into the fold. If we appraise success at both bandwidths simultaneously by averaging the two counts, as shown for each $b$-setting in Table 1's last two columns for each problem, we see that $a = +1$ and $a = -1$ have rather similar averaged success ratings, with $a = +1$ mildly superior to $a = -1$ under bandwidth target $b = .15$, but with the superiority order reversed under $b = .10$. This crossover interaction between parameters $a$ and $b$ can be explained by arguing similar to before that under $b = .10$ the indifference of curvature $a = -1$ to centering points within $\pm.10$ frees it for greater effort at bringing moderate outliers close to if not within the target band.

3. Increasing $a$ beyond +1 to +2 and +3 appears to degrade performance slightly, but scarcely enough to matter were there other grounds (I know of none) on which to prefer higher $a$. This suggests that ratio $h_{in}/h_{out}$ is less important than might have been anticipated.

4. Though Table 1 hints that higher powers $r = s = 4$ may not be quite so good as $r = s = 2$, at least when $a > 0$, the difference is small and unreliable. But because higher powers increase computation time even if only by a small percent, the salient observation is simply that nothing appears to be gained from these. (However, I have not explored this facet of parameter variation very thoroughly.)

---

[3] All POL entries in Table 1 were obtained under $p = .5$ with the exception of pattern $B$ with hyperplane-bandwidth target $b = .15$. The latter used $p = .9$, which worked best at the time. But more recently, after programming changes which should not have appreciably affected computational outcome, pattern $B$ with target $b = .15$ gets virtually these same results from $p = .5$ while $p = .9$ has become inferior in this case too.

We may provisionally conclude that there is no evident reason to bother in practice with power settings other than $r = s = 2$ (except perhaps for continued testing of the *Note 3* option called by negative $r$), nor with curvature $a$ greater than +1. But curvature variation within range $-1 \leq a \leq +1$ is an option worth retaining. What is intriguing about this is not that one setting of $a$ gives on the whole better hyperplane counts than does another, but that the *shape* of the factor-loading distribution within and near the target hyperplane-band is appreciably affected by it. What we might gain from controlling this, I have no idea. But it invites us to reflect on our preferences in the matter.

The largest uncertainty that remains for HYBALL applications is when to use SCAN rather than POL. Knowing that SCAN works best for complex problems, we would scarcely want to aschew it altogether. Yet to SCAN exclusively would be to ignore POL's vastly greater speed and frequent near-equivalence to SCAN. The most effective combination of these alternatives would seem to be POL for preliminary rotations, including tests of parameter variations, until no further improvement under POL seems possible. It is then appropriate to close with SCAN in hopes of kicking up the hyperplane count a percentage point or two. Be also advised, however, that for patterns whose latent simple structure is ambiguous, that is, which have two or more substantially different rotations with near-maximal hyperplane counts, which one HYBALL finds is much affected by starting position (Varimax vs. no Varimax) and how POL and SCAN are alternated thereafter. I have recently exploited the basis of this phenomenon (namely, convergence to optima that are merely local) for an important enhancement of HYBALL's prowess described in my forthcoming *The glory of suboptimal factor rotation: A silk-purse-into-sow's ear caper.*

*Other single-plane issues*

We still have some unfinished business in the question, raised previously but not pursued, whether unequal weights $\{c_i\}$ in Equation 1 are useful even when $\psi$ is bounded. All HYBALL's current variants give zero weight to extreme outliers (more precisely, to points that lie outside STEP's initial search window); however, this is mainly to short out computations that are essentially irrelevant, though it also makes negligible the risk of factor collapse. Yet there are more consequential ways to assign differential weights in Equation 1, indeed, far too many for systematic inventory. Even so, any such weighting scheme should be grounded on each point's position in the total pattern configuration. And for rotating $f_1$ in the $f_1/f_2$ plane, the prospect for $c_i$ that seems most promising is some measure of point $(a_{i1}, a_{i2})$'s prominence in this plane, say an increasing function of its distance from the origin. But points at a large angle to $f_1$ are mainly ones we want to de-emphasize in rotation of $f_1$ regardless

of their prominence; so it is really just the $|a_{i1}|$-component of $i$'s distance from origin — that is, this item's prospective salience as a marker of the $f_2$-hyperplane — that is our top candidate for $c$-weighting of hyperplane misfit. Accordingly, let us say that Equation 1 is *salience weighted* if each $c_i$ therein is an increasing function of $|a_{i1}|$. There are many such functions, and serious exploration of salience weighting would want to examine a number of them. But the class

$$c_i = |a_{i1}|^t \quad (t \geq 0), \tag{12}$$

proposed by Eber (1966) surely provides as much diversity in this respect as we are likely to want.[4] Uniform item weighting is Equation 12's limiting case wherein $t = 0$; while the differential weighting is gentle for $t$ under .5 but becomes quite severe if $t$ much exceeds 1. With $t = 2$, setting $a = 0$ in Equation 9 with large hyperplane bandwidth makes Equations 1/2 essentially the Direct Quartimin criterion.

Is salience weighting a good idea when fitting hyperplanes? Eber found it to be so, or at least we have his word (1966, p. 118) that Maxplane foundered without it. But Maxplane's all-or-none $\psi$-function yields an exceptional misfit measure that can incur large changes from very small axis shifts. In contrast, when $\psi(\varepsilon)$ increases continuously in $\varepsilon$, Equation 1 contains an implicit salience weighting even when the $c_i$ are all equal. For the change in $i$'s loading on $f_2$ produced by a small $f_1$-shift $w$ is proportional to $a_{i1}$, that is, $\tilde{a}_{i2} - a_{i2} = -wa_{i1}$, whence the more salient $i$ is on $f_1$ the larger is $i$'s contribution to change in $f_2$-hyperplane misfit under small shifts in $f_1$. So for continuous misfit functions, ones in which $\psi$ has Equation 9 form in particular, we can anticipate that there may well be little if anything to gain from additional explicit salience weighting.

And indeed, that has been my experience. Except for the outlier exclusions already mentioned, all results summarized in Table 1 were obtained under equal-weights setting $t = 0$ in Equation 12. But I have also tested $t = .5$ and $t = 1.0$ on these patterns under the main parameter variants of POL and SCAN, and in no case did $t > 0$ improve the targeted hyperplane counts more than trivially. Rather, $t > 0$ prevailingly resulted in some loss of hyperplane strength, with $t = 1.0$ mildly worse than $t = .5$; and for the most successful rotations of the more difficult patterns, replacement of equal weighting by $t = .5$ or $t = 1.0$ incurred hyperplane-count drops of several percentage points. Salience weighting will remain a HYBALL option to permit its further study; and of course we may find other weighting schemes superior to Equation 12. But provisionally, the word on salience weighting when fitting hyperplanes by misfit functions of the sort investigated here is: *Don't.*

---

[4] Eber further divides $|a_{i1}|$ by hyperplane-bandwidth $b$ before powering. But that is only a scaling adjustment having no effect on the weight ratios $(c_i/c_j)$.

Beyond some technicalities to be outlined without detail in the Appendix, I have no further wisdom on the theory and practice of analytic hyperplane optimization to share with you at this time. But we have scarcely put an end to the matter. In particular, a large apparent weakness in most extant approaches to this, including much of the present one, is dependence on an arbitrary choice of hyperplane-bandwidth $b$. Our intent for $b$, of course, is to discriminate pattern coefficients that differ appreciably from zero from ones that do not. But whereas what counts as *appreciable* should not be dictated by arbitrary scaling decisions, the numerical values of factor loadings $\{a_{ij}\}$ are strongly influenced by our choice of scale units for both the variables and the factors. (Conventional variance normalization is not so much a solution to this scaling problem as an evasion of it.) What we want is a logic for selecting an appropriate $b$ for each different combination of variable and factor, or, lacking that, at least an analytic criterion that prescribes an optimal $b$ for each factor axis without need for an arbitrary *deus-ex-machina* stipulation of this. But until such time as we can develop an operational theory of optimized $b$-settings, restricting misfit measures to Equation 11 form obviates much of this problem in theory even if we find other choices of $\psi$ preferable in practice.

## References


Cattell, R. B., & Muerle, J. L. (1960). The "Maxplane" program for factor rotation to oblique simple structure. *Educational and Psychological Measurement, 20*, 569-590.

Eber, H. W. (1966). Toward oblique simple structure: Maxplane. *Multivariate Behavioral Research, 1*, 112-125.

Harman, H. H. (1960). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31*, 313-323.

Katz, J. O., & Rohlf, F. J. (1974). Functionplane — a new approach to simple structure rotation. *Psychometrika, 39*, 37-51.

Katz, J. O., & Rohlf, F. J. (1975). Primary product functionplane: An oblique rotation to simple structure. *Multivariate Behavioral Research, 10*, 219-232.

Rozeboom, W. W. (1991). HYBALL: A method for subspace-constrained oblique factor rotation. *Multivariate Behavioral Research, 26*, 163-177.


## *Appendix*

Several technical issues which have arisen during my study of analytic hyperplane optimization are discussed rather extensively in a document available from me on request. This was originally prepared for appending to the present article as a prophylaxis against anticipated criticisms, albeit the points addressed hold intrinsic interest as well. However, simple mercy urges abstention from inflicting these esoterica upon others until such time as popular

demand compels their release. Even so, you should be apprised in outline of the details you have been spared. These fall under two main headings, (a) whether appraisal of factor-shift quality should explicitly include consideration of variance renormalization, and (b) the relation of Direct Quartimin to the present family of hyperplane-optimization criteria.

*Variance renormalization*

On the face of it, the present study of hyperplane appraisal is deeply flawed from its outset by insufficient concern for the full pattern effect even of shifting just one axis in a single factor plane, let alone the effects in other planes as well. For let $\gamma(\mathbf{a}_1, \mathbf{a}_2)$ be whatever function we have chosen to measure the inferiority of the $f_1$-hyperplane (i.e., loading configuration on $f_2$) in the $f_1/f_2$ plane when the column vectors of item loadings on $f_1$ and $f_2$ are respectively $\mathbf{a}_1$ and $\mathbf{a}_2$. My presumption that the hyperplane quality resulting from rotation

(A1) $$\tilde{f}_1 = f_1 + wf_2$$

of $f_1$ into $\tilde{f}_1$ with $f_2$ unchanged should be measured by $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2) = \gamma(\mathbf{a}_1, \mathbf{a}_2\text{-}w\mathbf{a}_1)$, is seemingly myopic on three counts: First, there is no reason why the coefficient of $f_1$ in this rotation must be unity. Secondly, Equation A1 fails to give $f_1$ the unit variance on which any right-minded factor analyst will insist (and which constraint I too accept even though variance normalization is merely a convention). And thirdly, this rotation affects other factor planes as well.

I commence rebuttal by pointing out that any rotation of $(f_1, f_2)$ into $(\hat{f}_1, f_2)$, where $\hat{f}_1$ is normalized and not collapsed into $f_2$, can always be treated as a two-stage operation that first replaces $f_1$ by $\tilde{f}_1$ defined by Equation A1 for some $w$, and then rescales $\tilde{f}_1$ into $\hat{f}_1 = \tilde{f}_1/s$ where $s^2$ is the variance of $\tilde{f}_1$. This variance normalization merely rescales the pattern column on the shifted factor as $\tilde{\mathbf{a}}_1 = s\mathbf{a}_1$ without further affecting the new pattern column on $f_2$. Even so, it seems to follow that the hyperplane quality resulting from this axis shift should really be measured by $\gamma(\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2) = \gamma(s\mathbf{a}_1, \mathbf{a}_2\text{-}w\mathbf{a}_2)$, where $s = (1 + 2wr_{12} + w^2)^{½}$ is a function of $w$ and the correlation $r_{12}$ between $f_1$ and $f_2$. Moreover, $\gamma(\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2)$ appraises only the rotated factor's hyperplane; whereas insomuch as the pattern loadings characterizing $f_2$'s hyperplane have also been shifted from $\mathbf{a}_1$ to $\tilde{\mathbf{a}}_1$, our assessment of this rotation's quality should include $\gamma(\tilde{\mathbf{a}}_2, \tilde{\mathbf{a}}_1)$ as well. Beyond that, alterations in pattern columns 1 and 2 also carry into the other planes for which one of these two factors is an axis; so $\gamma(\mathbf{a}_k, \tilde{\mathbf{a}}_1)$, $\gamma(\tilde{\mathbf{a}}_1, \mathbf{a}_k)$, $\gamma(\mathbf{a}_k, \tilde{\mathbf{a}}_2)$, and $\gamma(\tilde{\mathbf{a}}_2, \mathbf{a}_k)$ for $k = 3, 4, \ldots, m$ ought likewise to be heeded.

Apart from one complication, however, this argument can easily be rejected for all Equations 1/2 instantiations of $\gamma$. The complication concerns *configural* item-weighting under which the $c_i$ in Equation 1 are determined all or in part by

the full planar positions of the points whose loadings on $f_2$ we seek to minimize. (Salience-weighting scheme Equation 12 with $t > 0$ is the premier instance in point.) When $\gamma(\mathbf{a}_1, \mathbf{a}_2)$ does *not* incorporate configural weighting, however, it contains $\mathbf{a}_1$ only vacuously and can better be written $\gamma(\cdot, \mathbf{a}_2)$ to make clear that its first argument is irrelevant. In this case, insomuch as $\gamma(\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2)$ after variance renormalization is identical with the unnormalized $\gamma(\mathbf{a}_1, \tilde{\mathbf{a}}_2)$, it is pointless to reject the second in favor of the first. Even so, there is still the shift from $\gamma(\cdot, \mathbf{a}_1)$ to $\gamma(\cdot, \tilde{\mathbf{a}}_1)$ to consider. But letting that influence our solution for optimal repositioning of $f_1$ is a bad idea for two reasons: First, change in $s$ makes a difference for $\gamma(\cdot, s\mathbf{a}_1)$ merely by altering each $\mathbf{a}_1$-element's ratio to hyperplane-bandwidth parameter $b$ with no affect on these loadings' sizes in proportion to one another, and can be undone simply by opting to replace $b$ by $sb$ for the shifted factor's hyperplane while leaving it at $b$ for $f_2$. Because choice of $b$ is uncomfortably arbitrary in any case (see p. 193), we want to downplay its effect on rotation outcome and hence especially do not want what amounts to readjustment of $b$ for one factor to intrude into our appraisal of hyperplane quality for another. Moreover, it can also be shown that solving for $w$ to minimize a composite of $\gamma(\cdot, \mathbf{a}_2\text{-}w\mathbf{a}_1)$ and $\gamma(\cdot, s\mathbf{a}_1)$, rather than just the former, pulls the rotated axis away from its subjectively optimal position in idealized pattern configurations. (The demonstration's heart is that the dependence of $s$ on $r_{12}$ contaminates assessment of pure pattern quality.) Finally, it will be plain how vacuity of $\gamma$'s first argument makes pointless any concern for this shift's effects on factor planes other than its primary one.

If $\gamma(\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2)$ does encorporate configural item weighting, however, the analysis becomes much more intricate. For then, explicit renormalization of the shifted factor during form-Equation 1 rotation makes not only the $\varepsilon_i$-terms in Equation 1 a function of $w$ but the weights $c_i$ as well. There is no great difficulty in operationalizing such fluid weighting in search for minimal $E_\psi(w)$; it merely complicates the computation. But there is no anticipated benefit to repay that complication. Although details are strongly dependent on the particular configural weighting scheme at issue, it can be shown that permitting weights $\{c_i\}$ in Equation 1 to vary fluidly with $w$ tends to deflect solution for the $E_\psi$-minimizing $w$ from its subjectively optimal value unless the $c_i(w)$-functions are adjusted to minimize this bias — whereafter it should make little or no difference whether the $c_i(w)$ are allowed to vary during search for best $w$ rather than being held constant at their search-outset values $\{c_i(0)\}$.

*Conclusion*

Continuous factor-variance renormalization is a pleasant arena for conceptual games in the advanced theory of analytic hyperplane optimization. But it would be perverse to prefer this in practice, much less to insist on it.

*Direct Quartimin*

The desirability of comparing Direct Quartimin (henceforth DQ) to the present criteria for simple-structure rotation has been forcefully brought to my attention by referee comments saying, in gist, "Since connoisseurs have found DQ so satisfactory, who needs HYBALL?" The bottom-line rejoinder might be, "What's wrong with giving HYBALL a chance to compete?" But a more instructive response is to examine the similarities and differences here in some detail.

The DQ measure of hyperplane misfit in a single plane with factor loadings $\mathbf{a}_1$ and $\mathbf{a}_2$ on respective normalized factors $f_1$ and $f_2$ is

$$q(\mathbf{a}_1, \mathbf{a}_2) = \sum_{i=1}^{n} a_{i1}^2 a_{i2}^2.$$

When axis-shift $\tilde{f}_1 = f_1 + wf_2$ followed by variance renormalization $\hat{f}_1 = \tilde{f}_1/s$ rotates $f_1$ into $\hat{f}_1$, the value $\gamma(\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2)$ of $q$ for the shifted pattern as a function of rotation driver $w$ can be expressed in the form of Equations 1/2 as

(A2) $$E_q(w) = \sum_{i=1}^{n} c_i(w) \times \varepsilon_i(w),$$

(A3) $$\varepsilon_i(w) = (\tilde{a}_{i2}/b)^2 = [(a_{i2} - wa_{i1})/b]^2,$$

(A4) $$c_i(w) = b^2\tilde{a}_{i2}^2 = b^2a_{i2}^2(1 + 2wr_{12} + w^2),$$

with $r_{12}$, as before, the correlation between $f_1$ and $f_2$. Equation A2 is simply Equation 1 under the Identity specification of $\psi$, but with the weights $c_i$ that are fixed in Equation 1 now allowed to vary fluidly with $w$. Equation A3 is identical with Equation 2. And Equation A4 details how $c_i$ changes with $w$, albeit only its $a_{i2}^2$ component is salient here. Hyperplane-bandwidth parameter $b$ occurs vacuously in $E_q$, because it cancels out when Equation A3/A4 are inserted into Equation A2. But we show it to maximize the match between Equations 1/2 and Equations A2/A3/A4.

Equations A2/A3 show single-plane DQ to be the same as step-down regression except for a rather large difference in how weights $\{c_i\}$ are revised as $w$ varies in search of an optimum. DQ's weighting scheme is essentially a fluid version of Equation 12 with $t = 2$, which for salience weighting is quite severe. And unlike STEP's zero-weighting of points beyond its increasingly narrow search window, DQ's weights do nothing to lessen the focus of least-squares

regression on capturing the most extreme outliers. Consequently, although DQ lies within the family of simple-structure algorithms implemented by HYBALL — or, more precisely, would do so were HYBALL expanded to include fluid variance renormalization and to heed the secondary shift effects disdained earlier in this Appendix — its extreme salience weighting and even more its unrelentingly concave misfit curvature places it among the NOT RECOMMENDED variants. Direct Quartimin should excel at rotation to an independent-clusters pattern, but high-quality hyperplanes cannot be expected from it for data not well-approximating that ideal.