Rejoinder

David R. Hunter and Kenneth Lange

What's in a Name?

We thank the discussants for their insightful and substantive comments. In replying, we open with the least substantive issue—the name of the algorithm. We take the objections to "optimization transfer" well and find Professor Meng's call for a more attractive name compelling. Although we could not help but be amused by his suggestion of the "SM algorithm", we fear that the pun on sadomasochism might wear poorly over time. His suggestion also fails to meet a major objection of the other discussants. As both Groenen-Heiser and de Leeuw-Michailidis point out, a name like "Surrogate-Maximization" could apply equally well to Newton's method or steepest ascent with a line search.

We therefore propose the name "MM algorithm" as a compromise. Here MM stands for either Majorize-Minimize or Minorize-Maximize, depending on the context. Unlike SM or optimization transfer, this name tells us that the surrogate function is special because it either majorizes or minorizes the objective function. MM algorithm also echoes the names "majorization" and "iterative majorization" coined by earlier authors without risking confusion over the use of "majorization" in an entirely different branch of mathematics (Marshall and Olkin, 1979). Finally, MM algorithm emphasizes the affinity with the EM algorithm. Of course, MM has its own associations. As pointed out in the recent advertising campaign of M&M candy, the Roman numerals MM stand for the year 2000. Thus, MM not only accurately reflects the nature of the algorithm, but it also hints that the MM algorithm is the algorithm of the future. We could hardly ask for more in a new name.

Quadratic Matrix Trace Functions

We are grateful to Professor Kiers for discussing the optimization of a quadratic matrix trace function. In our view, this example fits easily within the scope of our paper.

Kiers' objective function $f(X) = \mathrm{tr}\,(BXCX^t)$ is concave, assuming the square matrices C and $-B$ are positive definite, and the goal is to minimize it subject to

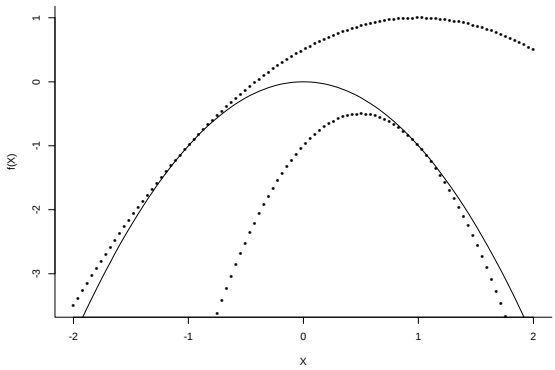

Figure 1: *The solid curve is the graph of the objective function $f(X)$ when X is a scalar and $C = -B = 1$. Because the second differential $d^2 f(X) = -2$ is constant, any tangent quadratic in this case furnishes either a minorizer or a majorizer of $f(X)$, according as its second differential is smaller than -2 (lower dotted curve) or greater than -2 (upper dotted curve). Note that requiring a majorizer to have positive second differential would make little sense here.*

constraints. Typically, of course, if constraints are absent, we are generally interested in *maximizing* a concave objective function. To achieve an algorithm which minimizes $f(X)$, we need a majorizing function. As illustrated by Figure 1 when X is a scalar, the spirit of the quadratic bound principle of Böhning and Lindsay (1988) may be invoked to produce either a majorizer or a minorizer of the given concave function.

Since $d^2 f(X) = 2(C \otimes B)$ is constant, the curvature of $f(X)$ is clearly bounded. We seek a majorizing quadratic which is concave and less curved than $f(X)$ in the sense that the second differential, say $2G$, of the majorizing quadratic satisfies $0 \succ 2G \succ 2(C \otimes B)$. The obvious choices $G = \lambda I$ and $G = \rho(C \otimes I)$, where λ and ρ are the largest eigenvalues of the negative definite matrices $C \otimes B$ and B, respectively, give the majorizing functions $h_2(X)$ and $h_3(X)$ displayed in Kiers' equations (5) and (11). Because these two choices of G satisfy $G \succ C \otimes B$, Kiers' inequality (9) must hold. We agree that the condition $G \succ 0$ should be dropped in this case since we are majorizing a concave rather than a convex function.

We are puzzled by Kiers' conclusion that "it has been seen that, in deriving majorizers for quadratic matrix functions, convexity need not play a role whatsoever." This view is hardly consistent with Meng's conjecture that all or nearly all MM algorithms are EM algorithms. If Meng is right, then convexity lies at the heart of all MM algorithms. In deference to Kiers, we certainly do not wish to suggest that

our paper provides an exhaustive list of the tools relevant to the creation of MM algorithms. In fact, we would be delighted to see the development of new examples, regardless of whether they depend on convexity or not. The beautiful example given by de Leeuw and Michailidis of "Dinkelbach majorization" does not appear to involve convexity, but then again it is not an MM algorithm.

A D-Optimal Design Problem

Professors de Leeuw and Michailidis raise the interesting issue of block relaxation algorithms. Here is a concrete example that combines the first step of the MM algorithm with block relaxation. In regression, D-optimal designs minimize the function $\ln\det(X^tX)^{-1}$, where X is the design matrix (Atkinson and Donev, 1996).

Since the log-determinant is a concave function on the space of positive definite matrices (Magnus and Neudecker, 1988), we derive the supporting hyperplane inequality

$$\ln\left|X^tX\right|^{-1} \;\leq\; \ln\left|X_n^tX_n\right|^{-1} + \mathrm{tr}\left\{X_n^tX_n\left[(X^tX)^{-1} - (X_n^tX_n)^{-1}\right]\right\} \tag{1}$$

using the differential

$$d\ln|M| \;=\; \mathrm{tr}\,(M^{-1}dM)$$

evaluated at $M = (X_n^tX_n)^{-1}$. Since equality holds in inequality (1) when $X = X_n$, the right side of that inequality gives the majorizing function

$$Q(X \mid X_n) \;=\; c(X_n) + \mathrm{tr}\left[X_n^tX_n(X^tX)^{-1}\right], \tag{2}$$

where $c(X_n)$ is an irrelevant constant. Thus, driving the trace in equation (2) downhill decreases the value of the objective function $\ln\left|X^tX\right|^{-1}$.

One possible way to exploit this majorization is to perturb one row of X at a time. In other words, we consider

$$X \;=\; \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix},$$

where $v_1, \ldots, v_m$ are row vectors. If we want to perturb v_1, then we write

$$\left(X^tX\right)^{-1} \;=\; \left(v_1^tv_1 + \cdots + v_m^tv_m\right)^{-1} \;=\; A^{-1} - \frac{A^{-1}v_1^tv_1A^{-1}}{1 + v_1A^{-1}v_1^t},$$

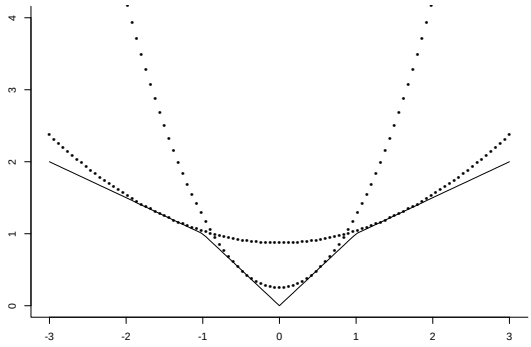

Figure 2: *The solid line is the graph of $g_p(u)$ corresponding to $p = 1/2$ and $d = 1$. The two dotted quadratic curves majorize $g_p(u)$ at $u^n = \pm 1/2$ and $u^n = \pm 3/2$.*

where $A = v_2^t v_2 + \cdots + v_m^t v_m$. Thus, the problem is to maximize, or at least increase, the value of

$$\text{tr}\left[\frac{X_n^t X_n A^{-1} v_1^t v_1 A^{-1}}{1 + v_1 A^{-1} v_1^t}\right] \quad = \quad \frac{v_1 A^{-1} X_n^t X_n A^{-1} v_1^t}{1 + v_1 A^{-1} v_1^t}$$

subject to relevant constraints. The process is then repeated for $v_2, \ldots, v_m$ in cyclic fashion. We have not yet implemented this promising algorithm.

GNC Algorithm

The GNC algorithm mentioned by Wu furnishes yet another example of an MM algorithm. For the sake of concreteness, suppose that

$$g_p(u) \quad = \quad \begin{cases} |u| & |u| \le d \\ p|u| + d(1-p) & |u| > d \end{cases}$$

for some positive threshold d and $p \in [0, 1]$. Despite the fact that $g_p(u)$ is neither convex nor concave for $p < 1$, we can clearly majorize it by a quadratic function. Figure 2 depicts the two cases where $u^n \neq 0$ lies inside and outside the central pit. Once we construct a quadratic majorizer $q_i(u)$ of $g_p(u)$ at $u_i^n = \theta_{i+1}^n - \theta_i^n$, we majorize $q_i(\theta_{i+1} - \theta_i)$ by the convex combination of quadratics

$$\frac{1}{2} q_i\Big[2\Big(\theta_{i+1} - \frac{\theta_{i+1}^n + \theta_i^n}{2}\Big)\Big] \quad + \quad \frac{1}{2} q_i\Big[-2\Big(\theta_i - \frac{\theta_{i+1}^n + \theta_i^n}{2}\Big)\Big] \tag{3}$$

with the parameters θ_{i+1} and θ_i separated. It follows that we can drive Wu's objective function $l_p(\theta)$ downhill by a simple linear update of each parameter θ_i holding the remaining parameters constant. It would be interesting to combine this tactic with a gradual lowering of p from 1 to 0.

Parameter Augmentation

It is also worth elaborating on Wu's comments about parameter augmentation. Consider the problem of estimating the location vector μ and the scale matrix Ω using m iid observations $x_1, \ldots, x_m$ from a multivariate t-distribution in R^p with ν degrees of freedom. If we let $\delta_i = (x_i - \mu)^t \Omega^{-1}(x_i - \mu)$ and $r_i = |\Omega|^a \delta_i$ with working parameter a, then the loglikelihood can be written as

$$\begin{aligned} L(\mu, \Omega) &= -\frac{m}{2} \ln |\Omega| - \frac{\nu + p}{2} \sum_{i=1}^{m} \ln \Big(1 + \frac{\delta_i}{\nu}\Big) \\ &= \frac{m[(\nu + p)a - 1]}{2} \ln |\Omega| - \frac{\nu + p}{2} \sum_{i=1}^{m} \ln \Big(|\Omega|^a + \frac{r_i}{\nu}\Big). \end{aligned}$$

Given the fact that $-\ln u$ is convex in u, it is trivial to show that up to a constant

$$Q(\mu, \Omega \mid \mu^n, \Omega^n) = \frac{m[(\nu + p)a - 1]}{2} \ln |\Omega| - \frac{\nu + p}{2} \sum_{i=1}^{m} w_i^n \Big(|\Omega|^a + \frac{r_i}{\nu}\Big)$$

minorizes $L(\mu, \Omega)$, where w_i^n is the weight $1/(|\Omega^n|^a + r_i^n/\nu)$. Regardless of the value of a, we accordingly update μ by

$$\mu^{n+1} = \frac{\sum_{i=1}^{m} w_i^n x_i}{\sum_{i=1}^{m} w_i^n}.$$

For the obvious choice $a = 0$ we update Ω by

$$\Omega^{n+1} = \frac{1}{n} \sum_{i=1}^{m} w_i^n (x_i - \mu^{n+1})(x_i - \mu^{n+1})^t.$$

The not so obvious choice $a = 1/(\nu + p)$ leads to the faster converging update

$$\Omega^{n+1} = \frac{\sum_{i=1}^{m} w_i^n (x_i - \mu^{n+1})(x_i - \mu^{n+1})^t}{\sum_{i=1}^{m} w_i^n}.$$

Although this is all spelled out by Meng and Van Dyk (1997), the MM perspective provides an easy derivation and eliminates some of the confusion of trying to relate parameter augmentation to missing data.

Optimality

We agree with Professor Kiers' suggestion that when several MM algorithms are possible for a problem, it would be good to have some basis for comparing them. One criterion proposed by Kiers is the closeness of the surrogate to the objective function. However, there is a danger in blindly adopting this criterion. The whole philosophy of the MM algorithm is to replace complex iterations with more but simpler iterations. Thus, simplicity—both computational and intuitive—should not be ignored.

Kiers' quadratic matrix function illustrates the issues well. If λ, the largest eigenvalue of $C \otimes B$, happens to be very close to zero, then the majorizer $h_2(X)$ of his equation (5) is not much different than the majorizer $h_1(X)$ of his equation (4). However, the computation needed to find λ and the extra term in $h_2(X)$ mean that the algorithm based on $h_2(X)$ will require more work per iteration than the $h_1(X)$ algorithm. Thus, if the iteration counts for the two algorithms are equal, the $h_1(X)$ version will enjoy an advantage not only in simplicity but also in speed, despite the fact that $h_1(X)$ is uniformly farther from the objective function than $h_2(X)$. We grant that in terms of computational time, $h_2(X)$ will probably outperform $h_1(X)$ in most practical problems and $h_3(X)$ might perform better still; yet $h_1(X)$ is certainly the easiest of the three to code, particularly if one does not have a greatest-eigenvalue-finding algorithm readily at hand.

At times, we may even achieve gains in simplicity by majorizing a majorizer. For example, in our discussion of the GNC algorithm, equation (3) shows how to construct a quadratic majorizer of a quadratic majorizer. Obviously, the majorizer of $q_i(u)$ is farther from the objective function than $q_i(u)$ itself; yet the gain in simplicity which equation (3) achieves by separating the parameters θ_i may be worthwhile. Erdoğan and Fessler (1999) use the idea of majorizing a majorizer to great effect in the field of transmission tomography, where the number of parameters is so large that directly solving for the minimum of a quadratic function is a practical impossibility unless the parameters are separated.

Oakes' Variance Decomposition

For the edification of the reader, it is worth emphasizing the simplicity of Meng's proof of Oakes' equation

$$d^2L(\theta) \;=\; d^{20}Q(\theta \mid \theta) + d^{11}Q(\theta \mid \theta).$$

If we consider θ to be a function $\theta(\phi)$ of ϕ and use the derivative notation of Dempster et al. (1977), then the equation

$$d^{10}H[\theta(\phi) \mid \phi] \;=\; 0$$

holds for the choices $\theta(\phi) = \phi$ and $H(\theta \mid \phi) = Q(\theta \mid \phi) - L(\theta)$. Differentiating this equation with respect to ϕ gives

$$d^{20}H[\theta(\phi) \mid \phi]\frac{d\theta}{d\phi} + d^{11}H[\theta(\phi) \mid \phi] \;=\; 0.$$

Substituting $\frac{d\theta}{d\phi} = I$ and

$$\begin{aligned} d^{20}H[\theta \mid \phi] &= d^{20}Q(\theta \mid \phi) - d^2L(\theta) \\ d^{11}H[\theta \mid \phi] &= d^{11}Q(\theta \mid \phi) \end{aligned}$$

and rearranging yields Oakes' result. The potential of this result in accelerating the MM and EM algorithms should not be overlooked.

Multidimensional Scaling

We are grateful to Groenen and Heiser for pointing out the discrepancies in speed between their tests of the various algorithms for MDS and ours. We looked at the code we used for our tests, adjusted it, and reran the tests. Suffice it to say that our new results closely resemble those of Groenen and Heiser as portrayed in their Figure 1. The biggest change from our Figure 3 is the fact that the de Leeuw-Heiser method, labeled SMACOF by Groenen and Heiser, is actually much faster than we originally reported. This difference is largely due to an unnecessary matrix times matrix multiplication at each iteration in our original code.

MMMCMC

In partial reply to Gelman, there are two ways of connecting the MM algorithm to Monte Carlo sampling. Suppose we majorize the logarithm $L(\theta)$ of a posterior density by $Q(\theta \mid \phi)$, which up to a known normalizing constant $c(\phi)$ is the logarithm of a probability density. We can implement the classical acceptance-rejection method of Monte-Carlo sampling by drawing θ randomly from $c(\phi)\exp[Q(\theta \mid \phi)]$ and U randomly from the uniform density on [0,1] and accepting θ provided

$$\ln U \;\leq\; L(\theta) - Q(\theta \mid \phi).$$

The fraction of sample points accepted is $c(\phi)$. To minimize the probability of rejection, one should choose ϕ to maximize $c(\phi)$. The proposal stage in this independent sampling method is particularly straightforward if $Q(\theta \mid \phi)$ is quadratic because a quadratic on the log scale corresponds to a normal distribution on the original scale.

On the other hand, we can generate correlated samples by letting ϕ equal the current sampled point θ^n. Regardless of whether $Q(\theta \mid \theta^n)$ majorizes or minorizes $L(\theta)$, we can implement Hastings-Metropolis sampling by using $c(\theta^n)\exp[Q(\theta \mid \theta^n)]$ as a proposal density and

$$\min\left\{\frac{e^{L(\theta^{n+1})}c(\theta^{n+1})e^{Q(\theta^n \mid \theta^{n+1})}}{e^{L(\theta^n)}c(\theta^n)e^{Q(\theta^{n+1} \mid \theta^n)}}, 1\right\}$$

as an acceptance probability. Since $L(\theta) = Q(\theta \mid \theta)$, this method may also be framed entirely in terms of the surrogate functions, which tend to be simpler to sample from than the log posterior. Although a quadratic is ideal if it closely approximates the log posterior, other log densities may work better on a given problem. In any case, the very techniques that we have suggested for constructing surrogate functions ought to prove valuable in Monte Carlo sampling as well.

MM versus EM

We find Meng's suggestion that any instance of MM might be expressible as an EM algorithm fascinating. Unfortunately, we do not yet have a cure for his "EM flu". We second his call for a search for an interesting MM algorithm that is not an EM algorithm or a definitive proof that no such example exists. However, even if the latter possibility turns out to be true, the MM framework is still more than "the old soup presented in a new, perhaps larger, bowl." As our data augmentation example demonstrates, the MM framework can provide simple derivations for algorithms which are unnecessarily complicated when viewed in the EM light. Even when a fairly natural EM algorithm exists for a given problem, it may be possible to construct a better MM algorithm. For instance, the early EM algorithm for transmission tomography discovered by Lange and Carson (1984) is decidedly inferior to the algorithm sketched in the present paper.

Closing

We hope that we and the discussants have convinced the reader of the potential of the MM algorithm. This is still much to be done, particularly in developing algorithms for high-dimensional problems. We thank the discussants for their thought-provoking critiques and their many prior contributions to this topic.

References

[1] Atkinson, A. C. and Donev, A. N. (1996). *Optimum Experimental Designs*. New York: Oxford.

[2] Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic approximation algorithms. *Annals of the Institute of Statistical Mathematics*, **40**: 641–663.

[3] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**: 1-38.

[4] Erdoğan, H. and Fessler, J. A. (1999). Monotonic algorithms for transmission tomography. *IEEE Transactions on Medical Imaging*, **18**: 801–814.

[5] Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, **8**: 306–316.

[6] Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

[7] Marshall, A. W. and Olkin, I (1979). *Inequalities: Theory of Majorization and its Applications*. San Diego: Academic Press.

[8] Meng, X.-L. and Van Dyk, D. A. (1997), The EM algorithm—an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 511–567.