

The MM Alternative to EM

Tong Tong Wu
Department of Epidemiology
and Biostatistics
University of Maryland
College Park, MD 20742

Kenneth Lange
Departments of Biomathematics,
Human Genetics, and Statistics
University of California
Los Angeles, CA 90095-1766
Phone: 310-206-8076
E-mail klange@ucla.edu

Research supported in part by USPHS grants
GM53275 and MH59490 to KL.

August 13, 2008

Abstract

The EM algorithm is a special case of a more general algorithm called the MM algorithm. Specific MM algorithms often have nothing to do with missing data. The first M step of an MM algorithm creates a surrogate function that is optimized in the second M step. In minimization, MM stands for majorize-minimize; in maximization, it stands for minorize-maximize. This two-step process always drives the objective function in the right direction. Construction of MM algorithms relies on recognizing and manipulating inequalities rather than calculating conditional expectations. This survey walks the reader through the construction of several specific MM algorithms. The potential of the MM algorithm in solving high-dimensional optimization and estimation problems is its most attractive feature. Our applications to random graph models, discriminant analysis, and image restoration showcase this ability.

Key Words: Iterative majorization, maximum likelihood, inequalities, penalization

1 Introduction

This survey paper tells a tale of two algorithms born in the same year. We celebrate the christening of the EM algorithm by Dempster et al. [17] for good reasons. The EM algorithm is one of the workhorses of computational statistics with literally thousands of applications. Its value was almost immediately recognized by the international statistics community. The more general MM algorithm languished in obscurity for years. Although in 1970 the numerical analysts Ortega and Rheinboldt [53] allude to the MM principle in the context of line search methods, the first statistical application occurs in two papers [12, 14] of de Leeuw and Heiser in 1977 on multidimensional scaling. One can argue that the unfortunate neglect of the de Leeuw and Heiser papers has retarded the growth of computational statistics. The purpose of the present paper is to draw attention to the MM algorithm and highlight some of its interesting applications.

Neither the EM nor the MM algorithm is a concrete algorithm. They are both principles for creating algorithms. The MM principle is based on the notion of (tangent) majorization. A function $g(\theta | \theta^n)$ is said to majorize a function $f(\theta)$ provided

$$\begin{aligned} f(\theta^n) &= g(\theta^n | \theta^n) \\ f(\theta) &\leq g(\theta | \theta^n), \quad \theta \neq \theta^n. \end{aligned} \tag{1}$$

In other words, the surface $\theta \mapsto g(\theta | \theta^n)$ lies above the surface $f(\theta)$ and is tangent to it at the point $\theta = \theta^n$. Here θ^n represents the current iterate in a search of the surface $f(\theta)$. The function $g(\theta | \theta^n)$ minorizes $f(\theta)$ if $-g(\theta | \theta^n)$ majorizes $-f(\theta)$. Readers should take heed that the term majorization is used in a different sense in the theory of convex functions [47].

In the minimization version of the MM algorithm, we minimize the sur-

rogate majorizing function $g(\theta | \theta^n)$ rather than the actual function $f(\theta)$. If θ^{n+1} denotes the minimum of the surrogate $g(\theta | \theta^n)$, then one can show that the MM procedure forces $f(\theta)$ downhill. Indeed, the relations

$$f(\theta^{n+1}) \leq g(\theta^{n+1} | \theta^n) \leq g(\theta^n | \theta^n) = f(\theta^n) \quad (2)$$

follow directly from the definition of θ^{n+1} and the majorization conditions (1). The descent property (2) lends the MM algorithm remarkable numerical stability. Strictly speaking, it depends only on decreasing the surrogate function $g(\theta | \theta^n)$, not on minimizing it. This fact has practical consequences when the minimum of $g(\theta | \theta^n)$ cannot be found exactly. In the maximization version of the MM algorithm, we maximize the surrogate minorizing function $g(\theta | \theta^n)$. Thus, the acronym MM does double duty, serving as an abbreviation of both pairs “majorize-minimize” and “minorize-maximize”. The earlier less memorable name “iterative majorization” for the MM algorithm unfortunately suggests that the principle is limited to minimization.

The EM algorithm is actually a special case of the MM algorithm. If $f(\theta)$ is the loglikelihood of the observed data, and $Q(\theta | \theta^n)$ is the function created in the E step, then the minorization

$$f(\theta) \geq Q(\theta | \theta^n) + f(\theta^n) - Q(\theta^n | \theta^n)$$

is the key to the EM algorithm. Maximizing $Q(\theta | \theta^n)$ with respect to θ drives $f(\theta)$ uphill. The proof of the EM minorization relies on the nonnegativity of the KullbackLeibler divergence of two conditional probability densities. The divergence inequality in turn depends on Jensen’s inequality and the concavity of the function $\ln x$ [29, 38].

In our opinion, the MM principle is easier to state and grasp than the EM principle. It requires neither a likelihood model nor a missing data framework. In some cases, existing EM algorithms can be derived more easily by

isolating a key majorization or minorization. In other cases, it is quicker and more transparent to postulate the complete data and calculate the conditional expectations required by the E step of the EM algorithm. Many problems involving the multivariate normal distribution fall into this latter category. Finally, EM and MM algorithms constructed for the same problem can differ. Our second example illustrates this point. Which algorithm is preferred is then a matter of reliability in finding the global optimum, ease of implementation, speed of convergence, and computational complexity.

This is not the first survey paper on the MM algorithm and probably will not be the last. The previous articles [3, 13, 24, 29, 40] state the general principle, sketch various methods of majorization, and present a variety of new and old applications. Prior to these survey papers, the MM principle surfaced in robust regression [25], correspondence analysis [23], the quadratic lower bound principle [8], alternating least squares applications [5, 33, 34, 59], medical imaging [18, 39], and convex programming [35]. Recent work has demonstrated the utility of MM algorithms in a broad range of statistical contexts, including quantile regression [27], survival analysis [28], nonnegative matrix factorization [19, 43, 44, 54], paired and multiple comparisons [26], variable selection [30], DNA sequence analysis [56], and discriminant analysis [21, 42].

The primary purpose of this paper is to present MM algorithms not featured in previous surveys. Some of these algorithms are novel, and some are minor variations on previous themes. Except for our first two examples in [Sections 2 and 3](#), it is unclear whether any of the algorithms can be derived from a missing data perspective. This fact alone distinguishes them from standard EM fare. In digesting the examples, readers should notice how the MM algorithm interdigitates with other algorithms such as block relaxation

and Newton's method. Classroom expositions of computational statistics leave the impression that different optimization algorithms act in isolation. In reality, some of the best algorithms are hybrids. The examples also stress penalized estimation and high-dimensional problems that challenge traditional algorithms such as scoring and Newton's method. Such problems are apt to dominate computational statistics and data mining for some time to come. The MM principle offers a foothold in the unforgiving terrain of large data sets and high dimensional models.

Two theoretical skills are necessary for constructing new MM algorithms. One is a good knowledge of statistical models. Another is proficiency with inequalities. Most inequalities are manifestations of convexity. The single richest source of minorizations is the supporting hyperplane inequality

$$f(x) \geq f(y) + df(y)(x - y)$$

satisfied by a convex function $f(x)$ at each point y of its domain. Here $df(y)$ is the row vector of partial derivatives of $f(x)$ at y .

The quadratic lower bound principle of Böhning and Lindsay [8] propels majorization when the objective function has bounded curvature. Let $d^2f(x)$ be the second differential (Hessian) of the objective function $f(x)$, and suppose B is a positive definite matrix such that $B - d^2f(x)$ is positive semidefinite for all arguments x . Then we have the majorization

$$\begin{aligned} f(x) &= f(y) + df(y)(x - y) + \frac{1}{2}(x - y)^t d^2f(z)(x - y) \\ &\leq f(y) + df(y)(x - y) + \frac{1}{2}(x - y)^t B(x - y), \end{aligned}$$

where z falls on the line segment between x and y . Minimization of the quadratic surrogate is straightforward. In the unconstrained case, it involves inversion of the matrix B , but this can be done once in contrast to the repeated matrix inversions of Newton's method. Other relevant majorizations

and minorizations will be mentioned as needed. Readers wondering where to start in brushing up on inequalities are urged to consult the elementary exposition [58]. The more advanced texts [10, 38] are also useful for statisticians.

Finally, let us stress that neither EM nor MM is a panacea. Optimization is as much art as science. There is no universal algorithm of choice, and a good deal of experimentation is often required to choose among EM, MM, scoring, Newton’s method, quasi-Newton methods, conjugate gradient, and other more exotic algorithms. The simplicity of MM algorithms usually argue in their favor. Balanced against this advantage is the sad fact that many MM algorithms exhibit excruciatingly slow rates of convergence. [Section 8](#) derives the theoretical criterion governing the rate of convergence of an MM algorithm. Fortunately, MM algorithms are readily amenable to acceleration. For the sake of brevity, we will omit a detailed development of acceleration and other important topics. Our discussion in [Section 9](#) will take these up and point out pertinent references.

2 Estimation with the Multivariate t

The multivariate t -distribution has density

$$f(x) = \frac{\Gamma\left[\frac{\nu+p}{2}\right]}{\Gamma\left(\frac{\nu}{2}\right) (\nu\pi)^{p/2} |\Omega|^{1/2} \left[1 + \frac{1}{\nu}(x - \mu)^t \Omega^{-1} (x - \mu)\right]^{(\nu+p)/2}}$$

for all $x \in \mathbb{R}^p$. Here μ is the mean vector, Ω is the positive definite scale matrix, and $\nu > 0$ is the degrees of freedom. Let x_1, \dots, x_m be a random sample from $f(x)$. To estimate μ and Ω for ν fixed, the well-known EM algorithm [41, 46] iterates according to

$$\mu^{n+1} = \frac{1}{s^n} \sum_{i=1}^m w_i^n x_i \tag{3}$$

$$\Omega^{n+1} = \frac{1}{m} \sum_{i=1}^m w_i^n (x_i - \mu^{n+1})(x_i - \mu^{n+1})^t, \quad (4)$$

where $s^n = \sum_{i=1}^m w_i^n$ is the sum of the case weights

$$w_i^n = \frac{\nu + p}{\nu + d_i^n}, \quad d_i^n = (x_i - \mu^n)^t (\Omega^n)^{-1} (x_i - \mu^n).$$

The derivation of the EM algorithm hinges on the representation of the t -density as a hidden mixture of multivariate normal densities.

Derivation of the same algorithm from the MM perspective ignores the missing data and exploits the concavity of the function $\ln x$. Thus, the supporting hyperplane inequality

$$-\ln x \geq -\ln y - \frac{x - y}{y}$$

implies the minorization

$$\begin{aligned} & -\frac{1}{2} \ln |\Omega| - \frac{\nu + p}{2} \ln [\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)] \\ \geq & -\frac{1}{2} \ln |\Omega| - \frac{\nu + p}{2} \left[\ln \frac{\nu + p}{w_i^n} + \frac{\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu) - (\nu + p)/w_i^n}{(\nu + p)/w_i^n} \right] \\ = & -\frac{1}{2} \ln |\Omega| - \frac{w_i^n}{2} [\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)] + c_i^n \end{aligned}$$

for case i , where c_i^n is a constant that depends on neither μ nor Ω . Summing over the different cases produces the overall surrogate. Derivation of the updates (3) and (4) reduces to standard manipulations with the multivariate normal [38].

Kent et al. [32] suggest an alternative algorithm that replaces the EM update (4) for Ω by

$$\Omega^{n+1} = \frac{1}{s^n} \sum_{i=1}^m w_i^n (x_i - \mu^{n+1})(x_i - \mu^{n+1})^t. \quad (5)$$

Meng and van Dyk [50] justify this modest amendment by expanding the parameter space to include a working parameter that is tweaked to produce

faster convergence. It is interesting that a trivial variation of our minorization produces the Kent et al. update. We simply combine the two log terms and minorize via

$$\begin{aligned}
& -\frac{1}{2} \ln |\Omega| - \frac{\nu + p}{2} \ln[\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)] \\
= & -\frac{\nu + p}{2} \ln\{|\Omega|^a [\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)]\} \\
\geq & -\frac{w_i^n}{2|\Omega^n|^a} \{|\Omega|^a [\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)]\} + c_i^n,
\end{aligned}$$

with working parameter $a = 1/(\nu + p)$.

For readers wanting the full story, we now indicate briefly how the second step of the MM algorithm is derived. This revolves around maximizing the surrogate function

$$-\sum_{i=1}^m w_i^n \{|\Omega|^a [\nu + (x_i - \mu)^t \Omega^{-1} (x_i - \mu)]\}$$

with respect to μ and Ω . Regardless of the value of Ω , one should choose μ as the weighted mean (3). If we let R be the square root of Ω^{n+1} as defined by equation (5) and substitute μ^{n+1} in the surrogate, then the refined surrogate function can be expressed

$$-s^n \{|\Omega|^a [\nu + \text{tr}(\Omega^{-1} R^2)]\} = -s^n \{|R^{-1} \Omega R^{-1}|^a [\nu + \text{tr}(R \Omega^{-1} R)]\} |R|^{2a}$$

To show that $\Omega = R^2$ minimizes the surrogate, let $\lambda_1, \dots, \lambda_p$ denote the eigenvalues of the positive definite matrix $R^{-1} \Omega R^{-1}$. This allows us to express the surrogate as a negative multiple of the function

$$h(\lambda) = \nu \prod_{j=1}^p \lambda_j^a + \prod_{j=1}^p \lambda_j^a \sum_{j=1}^p \lambda_j^{-1}.$$

The choice $\lambda = \mathbf{1}$ corresponds to $\Omega = R^2$ and yields the value $h(\mathbf{1}) = \nu + p$. The identity $\Omega = R^2$ can now be proved by showing that $\nu + p$ is a lower

bound for $h(\lambda)$. Setting $\lambda_j = e^{\theta_j}$, a simple rearrangement of the bounding inequality shows that it suffices to prove the alternative inequality

$$e^{-\frac{1}{\nu+p} \sum_{j=1}^p \theta_j} \leq \frac{\nu}{\nu+p} e^0 + \frac{1}{\nu+p} \sum_{j=1}^p e^{-\theta_j},$$

which is a direct consequence of the convexity of e^x .

3 Grouped Exponential Data

The EM algorithm for estimating the intensity of grouped exponential data is well known [17, 48, 51]. In this setting the complete data corresponds to a random sample x_1, \dots, x_m from an exponential density with intensity λ . The observed data conforms to a sequence of thresholds $t_1 < t_2 < \dots < t_m$. It is convenient to append the threshold $t_0 = 0$ to this list and to let c_i record the number of values that fall within the interval $(t_i, t_{i+1}]$. The exceptional count c_m represents the number of right-censored values. One can derive a novel MM algorithm by close examination of the loglikelihood

$$\begin{aligned} L(\lambda) &= c_0 \ln(1 - e^{-\lambda t_1}) + \sum_{i=1}^{m-1} c_i \ln(e^{-\lambda t_i} - e^{-\lambda t_{i+1}}) - c_m \lambda t_m \\ &= -\lambda \sum_{i=0}^{m-1} c_i t_{i+1} - c_m \lambda t_m + \sum_{i=0}^{m-1} c_i \ln(e^{\lambda d_i} - 1), \end{aligned}$$

where $d_i = t_{i+1} - t_i$.

The above partial linearization of the loglikelihood $L(\lambda)$ focuses our attention on the remaining nonlinear parts of $L(\lambda)$ determined by the function $f(\lambda) = \ln(e^{\lambda d} - 1)$. The derivatives

$$f'(\lambda) = \frac{e^{\lambda d} d}{e^{\lambda d} - 1}, \quad f''(\lambda) = -\frac{e^{\lambda d} d^2}{(e^{\lambda d} - 1)^2}$$

indicate that $f(\lambda)$ is increasing and concave. It is impossible to minorize $f(\lambda)$ by a linear function, so we turn to the quadratic lower bound principle.

Hence, in the second-order Taylor expansion

$$f(\lambda) = f(\lambda^n) + f'(\lambda^n)(\lambda - \lambda^n) + \frac{1}{2}f''(\mu)(\lambda - \lambda^n)^2,$$

with μ between λ and λ^n , we seek to bound $f''(\mu)$ from below. One can easily check that $f''(\mu)$ is increasing on $(0, \infty)$ and tends to $-\infty$ as μ approaches 0. To avoid this troublesome limit, we restrict λ to the interval $(\frac{1}{2}\lambda^n, \infty)$ and substitute $f''(\frac{1}{2}\lambda^n)$ for $f''(\mu)$. Minorizing the nonlinear part of $L(\lambda)$ term by term now gives a quadratic minorizer $q(\lambda)$ of $L(\lambda)$. Because the coefficient of λ^2 in $q(\lambda)$ is negative, the restricted maximum λ^{n+1} of $q(\lambda)$ occurs at the boundary $\frac{1}{2}\lambda^n$ whenever the unrestricted maximum occurs to the left of $\frac{1}{2}\lambda^n$. In symbols, the MM update reduces to

$$\lambda^{n+1} = \max \left\{ \frac{1}{2}\lambda^n, \lambda^n + \frac{\sum_{i=0}^{m-1} c_i (v_i^n - t_{i+1}) - c_m t_m}{\sum_{i=0}^{m-1} c_i w_i^n} \right\},$$

where

$$v_i^n = \frac{e^{\lambda^n d_i} d_i}{e^{\lambda^n d_i} - 1}, \quad w_i^n = \frac{e^{\lambda^n d_i/2} d_i^2 / 4}{(e^{\lambda^n d_i/2} - 1)^2}.$$

Table 1: Comparison of MM and EM on Grouped Exponential Data

n	MM Algorithm		EM Algorithm	
	λ^n	$L(\lambda^n)$	λ^n	$L(\lambda^n)$
0	1.00000	-3.00991	1.00000	-3.00991
1	0.50000	-1.75014	0.27082	-1.34637
2	0.25000	-1.32698	0.21113	-1.30591
3	0.18924	-1.30528	0.20102	-1.30443
4	0.19762	-1.30438	0.19904	-1.30437
5	0.19848	-1.30437	0.19864	-1.30437
6	0.19853	-1.30437	0.19856	-1.30437
7	0.19854	-1.30437	0.19854	-1.30437

Table 1 compares the MM algorithm and the traditional EM algorithm on the toy example of Meilijson [51]. Here we have $m = 3$ thresholds at 1, 3, and 10 and assign proportions 0.185, 0.266, 0.410, and 0.139 to the four ordinal groups. It is clear that the MM algorithm hits its lower bound on iterations 1 and 2. Although its local rate of convergence appears slightly better than that of the EM algorithm, the differences are minor. The purpose of this exercise is more to illustrate the quadratic lower bound principle in deriving MM algorithms.

4 Power Series Distributions

A family of discrete density functions $p_k(\theta)$ defined on $\{0, 1, \dots\}$ and indexed by a parameter $\theta > 0$ is said to be a power series family provided for all k

$$p_k(\theta) = \frac{c_k \theta^k}{q(\theta)}, \quad (6)$$

where $c_k \geq 0$ and $q(\theta) = \sum_{k=0}^{\infty} c_k \theta^k$ is the appropriate normalizing constant [55]. The binomial, negative binomial, Poisson, and logarithmic families are examples. Zero truncated versions of these families also qualify. Fisher scoring is the traditional approach to maximum likelihood estimation with a power series family. If x_1, \dots, x_m is a random sample from the discrete density (6), then the loglikelihood

$$L(\theta) = \sum_{i=1}^m x_i \ln \theta - m \ln q(\theta)$$

has score $s(\theta)$ and expected information $J(\theta)$

$$s(\theta) = \frac{1}{\theta} \sum_{i=1}^m x_i - \frac{mq'(\theta)}{q(\theta)}, \quad J(\theta) = \frac{m\sigma^2(\theta)}{\theta^2},$$

where $\sigma^2(\theta)$ is the variance of a single realization.

Functional iteration provides an alternative to scoring. It is clear that the maximum likelihood estimate $\hat{\theta}$ is a root of the equation

$$\bar{x} = \frac{\theta q'(\theta)}{q(\theta)}, \quad (7)$$

where \bar{x} is the sample mean. This result suggests the iteration scheme

$$\theta^{n+1} = \frac{\bar{x}q(\theta^n)}{q'(\theta^n)} = M(\theta^n) \quad (8)$$

and raises two obvious questions. First, is the algorithm (8) an MM algorithm? Second, is it likely to converge to $\hat{\theta}$ even in the absence of such a guarantee?. Local convergence hinges on the derivative condition $|M'(\hat{\theta})| < 1$. When this condition holds, the map $\theta^{n+1} = M(\theta^n)$ is locally contractive near the fixed point $\hat{\theta}$. It turns out that

$$M'(\hat{\theta}) = 1 - \frac{\sigma^2(\hat{\theta})}{\mu(\hat{\theta})},$$

where

$$\mu(\theta) = \frac{\theta q'(\theta)}{q(\theta)}$$

is the mean of a single realization X . Thus, convergence depends on the ratio of the variance to the mean. To prove these assertions it is helpful to differentiate $q(\theta)$. The first derivative delivers the mean and the second derivative the second factorial moment

$$E[X(X-1)] = \frac{\theta^2 q''(\theta)}{q(\theta)}.$$

If one substitutes these into the obvious expression for $M'(\hat{\theta})$ and invokes equality (7) at $\hat{\theta}$, then the moment form of $M'(\hat{\theta})$ emerges.

To address the question of whether functional iteration is an MM algorithm, we make the assumption that $q(\theta)$ is log-concave. This condition

holds for the binomial and Poisson distributions but not for the negative binomial and logarithmic distributions. The convexity of $-\ln q(\theta)$ entails the minorization,

$$\begin{aligned} L(\theta) &\geq \sum_{i=1}^m x_i \ln \theta - m \ln q(\theta^n) - m[\ln q(\theta^n)]'(\theta - \theta^n) \\ &= \sum_{i=1}^m x_i \ln \theta - m \ln q(\theta^n) - m \frac{q'(\theta^n)}{q(\theta^n)}(\theta - \theta^n). \end{aligned}$$

Setting the derivative of this surrogate function equal to 0 leads to the MM update (8). One can demonstrate that log-concavity implies $\sigma^2(\theta) \leq \mu(\theta)$. The local contraction condition $|M'(\hat{\theta})| < 1$ is consistent with the looser criterion $\sigma^2(\theta) \leq 2\mu(\theta)$. Thus, there is room for a viable local algorithm that fails to have the ascent property.

Table 2: Performance of the Algorithm (8) for Truncated Poisson Data

n	θ^n	$L(\theta^n)$	n	θ^n	$L(\theta^n)$
0	1.00000	-5.41325	7	1.59161	-4.34467
1	1.26424	-4.63379	8	1.59280	-4.34466
2	1.43509	-4.40703	9	1.59329	-4.34466
3	1.52381	-4.35635	10	1.59349	-4.34466
4	1.56424	-4.34670	11	1.59357	-4.34466
5	1.58151	-4.34501	12	1.59360	-4.34466
6	1.58867	-4.34472	13	1.59362	-4.34466

The truncated Poisson density has normalizing function $q(\theta) = e^\theta - 1$. The second derivative test shows that $q(\theta)$ is log-concave. Table 2 records the well-behaved MM iterates (8) for the choices $\bar{x} = 2$ and $m = 10$. The geometric density counting failures until a success has normalizing function $q(\theta) = (1 - \theta)^{-1}$, which is log-convex rather than log-concave. The iteration function is now $M(\theta) = \bar{x}(1 - \theta)$. Since $M'(\theta) = -\bar{x}$, the algorithm diverges

for $\bar{x} > 1$. Finally, the discrete logarithmic density has normalizing constant $q(\theta) = -\ln(1 - \theta)$, which is also log-convex rather than log-concave. The choices $\bar{x} = 2$ and $m = 10$ lead to the iterates in Table 3. Although the algorithm (8) converges for the logarithmic density, it cannot be an MM algorithm because the loglikelihood experiences a decline at its first iteration.

Table 3: Performance of the Algorithm (8) for Logarithmic Data

n	θ^n	$L(\theta^n)$	n	θ^n	$L(\theta^n)$
0	0.99000	-15.47280	9	0.71470	-8.98294
1	0.09210	-24.32767	10	0.71565	-8.98293
2	0.17545	-18.35307	11	0.71517	-8.98293
3	0.31814	-13.30624	12	0.71542	-8.98293
4	0.52221	-9.96349	13	0.71529	-8.98293
5	0.70578	-8.98560	14	0.71535	-8.98293
6	0.71991	-8.98355	15	0.71532	-8.98293
7	0.71291	-8.98310	16	0.71534	-8.98293
8	0.71655	-8.98297	17	0.71533	-8.98293

One of the morals of this example is that many natural algorithms only satisfy the descent or ascent property in special circumstances. This is not necessarily a disaster, but without such a guarantee, safeguards must usually be instituted to prevent iterates from going astray. Proof of the descent or ascent property almost always starts with majorization or minorization. Because so much of statistical inference revolves around loglikelihoods, log-convexity and log-concavity are possibly more important than ordinary convexity and concavity in constructing MM algorithms.

There are a variety of criteria that help in checking log-concavity. Besides the obvious second derivative test, one should keep in mind the closure properties of the collection of log-concave functions on a given domain [4, 10]. For example, the collection is closed under the formation of prod-

ucts and positive powers. Any positive concave function is log-concave. If $f(x) > \alpha \geq 0$ for all x , then $f(x) - \alpha$ is log-concave. In some cases, integration preserves log-concavity. If $f(x)$ is log-concave, then $\int_a^x f(y) dy$ and $\int_x^b f(y) dy$ are log-concave. When $f(x, y)$ is jointly log-concave in (x, y) , $\int f(x, y) dy$ is log-concave in x . As a special case, the convolution of two log-concave functions is log-concave. One of the more useful recent tests for log-concavity pertains to power series [1]. Suppose $f(x) = \sum_{k=0}^{\infty} a_k x^k$ has radius of convergence r around the origin. If the coefficients a_k are positive and the ratio $(k+1)a_{k+1}/a_k$ is decreasing in k , then $f(x)$ is log-concave on $(-r, r)$. This result also holds for finite series $f(x) = \sum_{k=0}^m a_k x^k$. In minorization, log-convexity plays the linearizing role of log-concavity. The closure properties of the set of log-convex functions are equally impressive [10].

5 A Random Graph Model

Random graphs provide interesting models of connectivity in genetics and internet node ranking. Here we consider the random graph model of Blitzstein et al. [6]. Their model assigns a nonnegative propensity p_i to each node i . An edge between nodes i and j then forms independently with probability $p_i p_j / (1 + p_i p_j)$. The most obvious statistical question in the model is how to estimate the p_i from data. Once this is done, we can rank nodes by their estimated propensities.

If E denotes the edge set of the graph, then the loglikelihood can be written as

$$L(p) = \sum_{\{i,j\} \in E} [\ln p_i + \ln p_j] - \sum_{\{i,j\}} \ln(1 + p_i p_j). \quad (9)$$

Here $\{i, j\}$ denotes a generic unordered pair. The logarithms $\ln(1 + p_i p_j)$ are

the bothersome terms in the loglikelihood. We will minorize each of these by exploiting the convexity of the function $-\ln(1+x)$. Application of the supporting hyperplane inequality yields

$$-\ln(1+p_i p_j) \geq -\ln(1+p_i^n p_j^n) - \frac{1}{1+p_i^n p_j^n} (p_i p_j - p_i^n p_j^n)$$

and eliminates the logarithm. Note that equality holds when $p_i = p_i^n$ for all i . This minorization is not quite good enough to separate parameters, however. Separation can be achieved by invoking the second minorizing inequality

$$-p_i p_j \geq -\frac{1}{2} \left(\frac{p_j^n}{p_i^n} p_i^2 + \frac{p_i^n}{p_j^n} p_j^2 \right).$$

Note again that equality holds when all $p_i = p_i^n$.

These considerations imply that up to a constant $L(p)$ is minorized by the function

$$g(p | p^n) = \sum_{\{i,j\} \in E} [\ln p_i + \ln p_j] - \sum_{\{i,j\}} \frac{1}{1+p_i^n p_j^n} \frac{1}{2} \left(\frac{p_j^n}{p_i^n} p_i^2 + \frac{p_i^n}{p_j^n} p_j^2 \right).$$

The fact that $g(p | p^n)$ separates parameters allows us to compute p_i^{n+1} by setting the derivative of $g(p | p^n)$ with respect to p_i equal to 0. Thus, we must solve

$$0 = \sum_{\{i,j\} \in E} \frac{1}{p_i} - \sum_{j \neq i} \frac{1}{1+p_i^n p_j^n} \frac{p_j^n}{p_i^n} p_i.$$

If $d_i = \sum_{\{i,j\} \in E} 1$ denotes the degree of node i , then the positive square root

$$p_i^{n+1} = \left[\frac{p_i^n d_i}{\sum_{j \neq i} \frac{p_j^n}{1+p_i^n p_j^n}} \right]^{1/2} \quad (10)$$

is the pertinent solution. Blitzstein et al. [6] derive a different and possibly more effective algorithm by a contraction mapping argument.

The MM update (10) is not particularly intuitive, but it does have the virtue of algebraic simplicity. When $d_i = 0$, it also makes the sensible choice $p_i^{n+1} = 0$. As a check on our derivation, observe that a stationary point of the loglikelihood satisfies

$$0 = \frac{d_i}{p_i} - \sum_{j \neq i} \frac{p_j}{1 + p_i p_j},$$

which is just a rearranged version of the update (10) with iteration superscripts suppressed.

The MM algorithm just derived carries with it certain guarantees. It is certain to increase the loglikelihood at every iteration, and if its maximum value is attained at a unique point, then it will also converge to that point. It is straightforward to prove that the loglikelihood is concave under the reparameterization $p_i = e^{-q_i}$. The requirement of two successive minorizations in our derivation gives us pause because if minorization is not tight, then convergence is slow. On the other hand, if the number of nodes is large, then competing algorithms such as Newton’s method entail large matrix inversions and are very expensive.

As a test case for the MM algorithm, we generated a random graph on $m = 10,000$ nodes with a propensity p_i for node i of $(i - \frac{1}{2})/m$. To derive appropriate starting values for the propensities, we estimated a common background propensity q by setting $q^2/(1 + q^2)$ equal to the ratio of observed edges to possible edges and solving for q . This background propensity was then used to estimate each p_i by setting $p_i q/(1 + p_i q)$ equal to d_i/m and solving for p_i . Table 4 displays the components p_0^n , $p_{m/2}^n$, and p_m^n of the parameter vector p^n at iteration n . The loglikelihood actually fails the ascent test in the last iteration because its rightmost digits are beyond machine precision. Despite this minor flaw, the algorithm performs impressively on

Table 4: Convergence of the MM Random Graph Algorithm

n	p_0^n	$p_{m/2}^n$	p_m^n	$L(p^n)$
0	0.00100	0.48240	0.95613	-40572252.7109
1	0.00000	0.48281	0.97251	-40565250.8333
2	0.00000	0.48220	0.98274	-40562587.5350
3	0.00000	0.48151	0.98950	-40561497.1411
4	0.00000	0.48093	0.99408	-40561038.9534
5	0.00000	0.48050	0.99720	-40560843.3998
10	0.00000	0.47963	1.00299	-40560695.6515
15	0.00000	0.47950	1.00387	-40560693.1245
20	0.00000	0.47948	1.00400	-40560693.0770
25	0.00000	0.47948	1.00403	-40560693.0761
30	0.00000	0.47948	1.00403	-40560693.0761
35	0.00000	0.47948	1.00403	-40560693.0764

this relatively large and decidedly non-sparse problem. As an indication of the quality of the final estimate \hat{p} , the maximum error $\max_i |(i - \frac{1}{2})/m - \hat{p}_i|$ was 0.0825 and the average absolute error $\frac{1}{m} \sum_i |(i - \frac{1}{2})/m - \hat{p}_i|$ was 0.0104.

6 Discriminant Analysis

Discriminant analysis is another attractive application. In discriminant analysis with two categories, each case i is characterized by a feature vector z_i and a category membership indicator y_i taking the values -1 or 1 . In the machine learning approach to discriminant analysis [57, 61], the hinge loss function $[1 - y_i(\alpha + z_i^t \beta)]_+$ plays a prominent role. Here $(u)_+$ is shorthand for the convex function $\max\{u, 0\}$. Just as in ordinary regression, we can penalize the overall loss

$$g(\theta) = \sum_{i=1}^n [1 - y_i(\alpha + z_i^t \beta)]_+$$

by imposing a lasso or ridge penalty [22]. Note that the linear regression function $h_i(\theta) = \alpha + z_i^t \beta$ predicts either -1 or 1 . If $y_i = 1$ and $h_i(\theta)$ over-predicts in the sense that $h_i(\theta) > 1$, then there is no loss. Similarly, if $y_i = -1$ and $h_i(\theta)$ under-predicts in the sense that $h_i(\theta) < -1$, then there is no loss.

Most strategies for estimating θ pass to the dual of the original minimization problem. A simpler strategy is to majorize each contribution to the loss by a quadratic and minimize the surrogate loss plus penalty. A little calculus [21] shows that $(u)_+$ is majorized at $u^n \neq 0$ by the quadratic

$$q(u \mid u^n) = \frac{1}{4|u^n|} (u + |u^n|)^2. \quad (11)$$

In fact, this is the best quadratic majorizer [16]. To avoid the singularity at 0 , we recommend replacing $q(u \mid u^n)$ by

$$r(u \mid u^n) = \frac{1}{4|u^n| + \epsilon} (u + |u^n|)^2.$$

In double precision, a good choice of ϵ is 10^{-5} . If we impose a ridge penalty, then the majorization (11) leads to a pure MM algorithm exploiting weighted least squares.

If the number of predictors is large, then the matrix inversions entailed in updating all parameters simultaneously become burdensome. Coordinate descent offers a viable alternative because it updates a single parameter at a time. The large number of iterations until convergence required by coordinate descent is often outweighed by the extreme simplicity of each parameter update. Quadratic majorization of the hinge losses keeps the updates simple and guarantees a reduction in the objective function. The decisions to use a lasso or ridge penalty and apply pure MM or coordinate descent with majorization will be dictated in practical problems by considerations of model selection and the number of potential predictors.

In discriminant analysis with more than two categories, it is convenient to pass to ϵ -insensitive loss and multiple linear regression. Our recently introduced method of vertex discriminant analysis (VDA) [42] operates in this fashion and relies on an MM algorithm. If there are $k + 1$ categories and p predictors, the basic idea is situate the class indicators at the vertices of a regular simplex in \mathbb{R}^k and minimize the criterion

$$R(A, b) = \frac{1}{n} \sum_{i=1}^n \|y_i - Az_i - b\|_\epsilon + \lambda \sum_{j=1}^k \|a_j\|^2, \quad (12)$$

where y_i is the vertex assigned to case i , a_j^t is the j th row of a $k \times p$ matrix A of regression coefficients, b is a $k \times 1$ column vector of intercepts, and

$$\|v\|_\epsilon = \max\{\|v\| - \epsilon, 0\} \quad (13)$$

is ϵ -insensitive Euclidean distance. Once A and b are estimated, we can assign a new case to the closest vertex, and hence category. One can design a quadratic surrogate by application of the Cauchy-Schwarz inequality and minimize the surrogate by solving k coordinated least squares problems. The combination of a parsimonious loss function and an efficient MM algorithm make VDA one of the most effective discriminant analysis methods tested [42].

As a comparison of hinge-loss discriminant analysis versus VDA, we now consider four typical examples from the UCI machine learning repository [2]. All four examples involve just two categories. For each data set, Table 5 lists the numbers of cases, features, and iterations until convergence, as well as the training error rates and the computing times in seconds under both hinge loss and ϵ -insensitive loss. For VDA we set $\epsilon = 0.9999$, just below the recommended cutoff of $\sqrt{(2k + 2)/k}/2 = 1$ for $k + 1 = 2$ categories. The cutoff is the largest ϵ avoiding overlap of the ϵ -insensitive spheres around

each vertex of the regular simplex. We chose the value 10^{-2} for the tuning parameter λ in all four examples. Our previous numerical experience shows that VDA is relatively insensitive to the choice of λ . Inspection of the training errors suggest that the two methods have similar accuracy. To our surprise, VDA is considerably faster.

Table 5: Empirical Examples from UCI Machine Learning Repository

Data Set (Cases,Features)	Hinge Loss			VDA		
	Iters	Error	Time	Iters	Error	Time
Diabetes (768,8)	44	0.2266	0.063	11	0.2240	0.015
SPECT (80,22)	326	0.2000	0.359	7	0.1750	0.000
Tic-tac-toe (958,9)	274	0.0167	0.578	26	0.0167	0.062
Ionosphere (351,33)	483	0.0513	2.984	42	0.0570	0.266

7 Image Restoration and Inpainting

The MM algorithm is also employed in image deconvolution [7, 45]. Suppose a photograph is divided into pixels and y_{ij} is the digitized intensity for pixel (i, j) . Some of the y_{ij} are missing or corrupted. Smoothing pixel values can give a visually improved image. Correction of pixels subject to minor corruption is termed denoising; correction of missing or grossly distorted values is termed inpainting. Let S be the set of pixels with acceptable values. We can restore the photograph by minimizing the criterion

$$\sum_{(i,j) \in S} (y_{ij} - \mu_{ij})^2 + \lambda \sum_i \sum_j \sum_{(k,l) \in N_{ij}} \|\mu_{ij} - \mu_{kl}\|_{\text{TV}},$$

where N_{ij} denotes the pixels neighboring pixel (i, j) , $\|x\|_{\text{TV}} = \sqrt{x^2 + \epsilon}$ is the total variation norm with $\epsilon > 0$ small, and $\lambda > 0$ is a tuning constant. Let μ_{ij}^n be the current iterate. The total variation penalties are majorized

using

$$\|x\|_{\text{TV}} \leq \|x^n\|_{\text{TV}} + \frac{1}{2\|x^n\|_{\text{TV}}} [x^2 - (x^n)^2]$$

based on the concavity of the function $\sqrt{t + \epsilon}$. These maneuvers construct a simple surrogate function expressible as a weighted sum of squares. Other roughness penalties are possible. For instance, the scaled sum of squares $\lambda \sum_i \sum_j \sum_{(k,l) \in N_{ij}} (\mu_{ij} - \mu_{kl})^2$ is plausible. Unfortunately, this choice tends to deter the formation of image edges. The total variation alternative is preferred in practice because it is gentler while remaining continuously differentiable.

If the pixels are defined on a rectangular grid, then we can divide them into two blocks in a checkerboard fashion, with the red checkerboard squares falling into one block and the black checkerboard squares into the other block. Within a block, the least squares problems generated by the surrogate function are parameter separated and hence trivial to solve. Thus, it makes sense to alternate the updates of the blocks. Within a block we update μ_{ij} via

$$\mu_{ij}^{n+1} = \frac{2y_{ij} + \lambda \sum_{(k,l) \in N_{ij}} \frac{\mu_{kl}^n}{\|\mu_{ij}^n - \mu_{kl}^n\|_{\text{TV}}}}{2 + \lambda \sum_{(k,l) \in N_{ij}} \frac{1}{\|\mu_{ij}^n - \mu_{kl}^n\|_{\text{TV}}}}$$

for $(i, j) \in S$ or via

$$\mu_{ij}^{n+1} = \frac{\sum_{(k,l) \in N_{ij}} \frac{\mu_{kl}^n}{\|\mu_{ij}^n - \mu_{kl}^n\|_{\text{TV}}}}{\sum_{(k,l) \in N_{ij}} \frac{1}{\|\mu_{ij}^n - \mu_{kl}^n\|_{\text{TV}}}}$$

for $(i, j) \notin S$. Here each interior pixel (i, j) has four neighbors. If the singularity constant ϵ is too small or if the tuning λ is too large, then small residuals generate very large weights. When this pitfall is avoided, the described algorithm is apt to be superior to the fused lasso algorithm of Friedman et al. [20].



Figure 1: Restoration of the Lenna Photograph. Top row left: the original image; top row right: image corrupted by Gaussian noise (mean 0 and standard deviation 10) and a scratch; second row left: restored image with $\lambda = 10$; second row right: restored image with $\lambda = 15$; third row left: restored image with $\lambda = 20$; third row right: restored image with $\lambda = 25$. The same value $\epsilon = 1$ is used throughout.

We applied the total variation algorithm to the standard image of the model Lenna. Figure 1 shows the original 256×256 image with pixel values digitized on a gray scale from 0 to 255. To the right of the original image is a version corrupted by Gaussian noise (mean 0 and standard deviation 10) and a scratch on the shoulder. The images are restored with λ values of 10, 15, 20, and 25 and an ϵ value of 1. Although we tend to prefer the restoration on the right in the second row, this is a matter of judgment. Variations in λ clearly control the balance between image smoothness and loss of detail.

8 Local Convergence of MM Algorithms

Many MM and EM algorithms exhibit a slow rate of convergence. How can one predict the speed of convergence of an MM algorithm and choose between competing algorithms? Consider an MM map $M(\theta)$ for minimizing the objective function $f(\theta)$ via the surrogate function $g(\theta | \theta^n)$. According to a theorem of Ostrowski [52], the local rate of convergence of the sequence $\theta^{n+1} = M(\theta^n)$ is determined by the spectral radius ρ of the differential $dM(\theta^\infty)$ at the minimum point θ^∞ of $f(\theta)$. Well-known calculations [17, 36] demonstrate that

$$dM(\theta^\infty) = I - d^2g(\theta^\infty | \theta^\infty)^{-1}d^2f(\theta^\infty).$$

Hence, the eigenvalue equation $dM(\theta^\infty)v = \lambda v$ can be rewritten as

$$d^2g(\theta^\infty | \theta^\infty)v - d^2f(\theta^\infty)v = \lambda d^2g(\theta^\infty | \theta^\infty)v.$$

Taking the inner product of this with v , we can solve for λ in the form

$$\lambda = 1 - \frac{v^t d^2f(\theta^\infty)v}{v^t d^2g(\theta^\infty | \theta^\infty)v}.$$

Extension of this line of reasoning shows that the spectral radius satisfies

$$\rho = 1 - \min_{v \neq 0} \frac{v^t d^2f(\theta^\infty)v}{v^t d^2g(\theta^\infty | \theta^\infty)v}.$$

Thus, the rate of convergence of the MM iterates is determined by how well $d^2g(\theta^\infty | \theta^\infty)$ approximates $d^2f(\theta^\infty)$. In practice, the surrogate function $g(\theta | \theta^n)$ should hug $f(\theta)$ as tightly as possible for θ close to θ^n .

Meng and van Dyk [50] use this Rayleigh quotient characterization of the spectral radius to prove that the Kent et al. multivariate t algorithm is faster than the original multivariate t algorithm. In essence, they show that the second differential $d^2g(\theta | \theta)$ is uniformly more positive definite for the alternative algorithm. de Leeuw and Lange [16] make substantial progress in designing optimal quadratic surrogates. For most other MM algorithms, however, such theoretical calculations are too hard to carry out, and one must rely on numerical experimentation to determine the rate of convergence. The uncertainties about rates of convergence are reminiscent of the uncertainties surrounding MCMC methods. This should not deter us from constructing MM algorithms. On large scale problems, many traditional algorithms are simply infeasible. If we can construct a MM algorithm, then there is always the chance of accelerating it. We take up this topic briefly in the discussion. Finally, let us stress that the number of iterations until convergence is not the sole determinant of algorithm speed. Computational complexity per iteration also comes into play. On this basis, a standard MM algorithm for transmission tomography is superior to a plausible but different EM algorithm [38].

9 Discussion

Perhaps, the best evidence of the pervasive influence of the EM algorithm is the sheer number of citations garnered by the Dempster et al. paper [17]. As of April 2008, Google Scholar lists 11,232 citations. By contrast, Google Scholar lists 58 citations for the de Leeuw paper [12] and 47 citations for

the de Leeuw and Heiser paper [14]. If our contention about the relative importance of the EM and MM algorithms is true, how can one account for this disparity? Several reasons come to mind. One is the venue of publication. The *Journal of the Royal Statistical Society, Series B*, is one of the most widely read journals in statistics. The de Leeuw and Heiser papers are buried in a hard to access conference proceedings. Another reason is the prestige of the authors. Four of the five authors of the three papers, Nan Laird, Donald Rubin, Jan de Leeuw, and Willem Heiser, were quite junior in 1977. On the other hand, Arthur Dempster was a major figure in statistics and well established at Harvard, the most famous American university. Besides these extrinsic differences, the papers have intrinsic differences that account for the better reception of the Dempster et al. paper. Its most striking advantage is the breadth of its subject matter. Dempster et al. were able to unify different branches of computational statistics under the banner of a clearly enunciated general principle. de Leeuw and Heiser stuck to multi-dimensional scaling. Their work and extensions are well summarized by Borg and Groenen [9].

The EM algorithm immediately appealed to the stochastic intuition of statisticians, who are good at calculating the conditional expectations required by the E step. The MM algorithm relies on inequalities and does not play as well to the strengths of statisticians. Partly for this reason the MM algorithm had difficulty breaking out of the vast but placid backwater of social science applications where it started. It remained sequestered there for years, nurtured by several highly productive Dutch statisticians with less clout than their American and British colleagues.

Our emphasis on concrete applications neglects some issues of considerable theoretical and practical importance. The most prominent of these

are global convergence analysis, computation of asymptotic standard errors, acceleration, and approximate solution of the optimization step (second M) of the MM algorithm. Let us address each of these in turn.

Virtually all of the convergence results announced by Dempster et al. [17] and corrected by Wu [63] and Boyles [11] carry over to the MM algorithm. The known theory, both local and global, is summarized in the references [38, 60]. As anticipated, the best results hold in the presence of convexity or concavity. The SEM algorithm of Meng and Rubin [49] for computation of asymptotic standard errors also carries over to the MM algorithm [26]. Numerical differentiation of the score function is a viable competitor, particularly if the score can be evaluated analytically. The simplest form of acceleration is step doubling [15, 39]. This maneuver replaces the point delivered by an algorithm map $\theta^{n+1} = M(\theta^n)$ by the new point $\theta^n + 2[M(\theta^n) - \theta^n]$. Step doubling usually halves the number of iterations until convergence in an MM algorithm. More effective forms of acceleration are possible using matrix polynomial extrapolation [62] and quasi-Newton and conjugate gradient elaborations of the MM algorithm [31, 37]. Finally, if the optimization step of an MM algorithm cannot be accomplished analytically, it is possible to fall back on the MM gradient algorithm [29, 36]. Here one substitutes one step of Newton's method for full optimization of the surrogate function $g(\theta \mid \theta^n)$ with respect to θ . Fortunately, this approximate algorithm has exactly the same rate of convergence as the original MM algorithm. It also preserves the descent or ascent property of the MM algorithm close to the optimal point.

The reader may be left wondering whether EM or MM provides a clearer path to the derivation of new algorithms. In the absence of a likelihood function, it is difficult for EM to work its magic. Even so, criteria such as

least squares can involve hidden likelihoods. Perhaps, the best reply is that we are asking the wrong question. After all, one man's mathematical meat is often another man's mathematical poison. A better question is whether MM broadens the possibilities for devising new algorithms. In our view, the answer to the second question is a resounding yes. Our last four examples illustrate this point. Of course, it may be possible to derive one or more of these algorithms from the EM perspective, but we have not been clever enough to do so.

In highlighting the more general MM algorithm, we intend no disrespect to the pioneers of the EM algorithm. If the fog of obscurity lifts from the MM algorithm, it will not detract from their achievements. It may, however, propel the ambitious plans for data mining underway in the 21st century. Even with the expected advances in computer hardware, the statistics community still needs to concentrate on effective algorithms. The MM principle is poised to claim a share of the credit in this enterprise. Statisticians with a numerical bent are well advised to add it to their toolkits.

References

- [1] Anderson GD, Vamanamurthy MK, Vuorinen M (2007) Generalized convexity and inequalities. <http://arxiv.org/abs/math/0701262>, arXiv
- [2] Asuncion A, Newman DJ (2007). UCI Machine Learning Repository [<http://www.ics.uci.edu/mlearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
- [3] Becker MP, Yang I, Lange K (1997) EM algorithms without missing data. *Stat Methods Med Res* 6:38–54

- [4] Bergstrom TC, Bagnoli M (2005) Log-concave probability and its applications. *Economic Theory* 26:445–469
- [5] Bijleveld CCJH, de Leeuw J (1991) Fitting longitudinal reduced-rank regression models by alternating least squares. *Psychometrika* 56:433–447
- [6] Blitzstein J, Chatterjee S, Diaconis P (2008). A new algorithm for high dimensional maximum likelihood estimation. (in preparation)
- [7] Bioucas-Dias JM, Figueiredo MAT, Oliveira JP (2006) Total variation-based image deconvolution: a majorization-minimization approach. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings, II* 861–864
- [8] Böhning D, Lindsay BG (1988) Monotonicity of quadratic approximation algorithms. *Annals of the Institute of Statistical Mathematics* 40:641–663
- [9] Borg I, Groenen P (1997) *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, New York
- [10] Boyd S, Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press, Cambridge
- [11] Boyles RA (1983) On the convergence of the EM algorithm. *J Roy Stat Soc B* 45:47–50
- [12] de Leeuw J (1977) Applications of convex analysis to multidimensional scaling. *Recent Developments in Statistics*, edited by Barra JR, Brodeau F, Romie G, Van Cutsem B, Elsevier/North-Holland, Amsterdam, pp 133–145

- [13] de Leeuw J (1994) Block relaxation algorithms in statistics. *Information Systems and Data Analysis*, (edited by Bock HH, Lenski W, Richter MM), Springer-Verlag, Berlin pp 308–325
- [14] de Leeuw J, Heiser WJ (1977) Convergence of correction matrix algorithms for multidimensional scaling. *Geometric Representations of Relational Data*, edited by Lingoes JC, Roskam E, Borg I). Mathesis Press, Ann Arbor, MI pp 735–752
- [15] de Leeuw J, Heiser WJ (1980) Multidimensional scaling with restrictions on the configuration. *Multivariate Analysis, Volume V*, edited by Krishnaiah PR, North-Holland, Amsterdam, pp 501–522
- [16] de Leeuw J, Lange K (2008) Sharp quadratic majorization in one dimension.
- [17] Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc B* 39:1–38
- [18] De Pierro AR (1995) A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans Med Imaging* 14:132–137
- [19] Eldén L (2007) *Matrix Methods in Data Mining and Pattern Recognition*. SIAM, Philadelphia
- [20] Friedman J, Hastie T, Tibshirani R (2007) Pathwise coordinate optimization.
- [21] Groenen PJF, Nalbantov G, Bioch JC (2007) Nonlinear support vector machines through iterative majorization and I-splines. *Studies in Clas-*

- sification, Data Analysis, and Knowledge Organization*, edited by Lenz HJ, Decker R, Springer-Verlag, Heidelberg-Berlin, pp 149–161
- [22] Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York
- [23] Heiser WJ (1987), Correspondence analysis with least absolute residuals. *Comput Stat Data Analysis* 5:337–356
- [24] Heiser WJ (1995) Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. *Recent Advances in Descriptive Multivariate Analysis*. (edited by Krzanowski WJ), Clarendon Press, Oxford, pp 157–189
- [25] Huber PJ (1981) *Robust Statistics*. Wiley, New York
- [26] Hunter DR (2004), MM algorithms for generalized Bradley-Terry models. *Annals Stat*
- [27] Hunter DR, Lange K (2000) Quantile regression via an MM algorithm. *J Comput Graphical Stat* 9:60–77
- [28] Hunter DR, Lange K (2002) Computing estimates in the proportional odds model. *Ann Inst Stat Math* 54:155–168
- [29] Hunter DR, Lange K (2004) A tutorial on MM algorithms. *Amer Statistician* 58:30–37
- [30] Hunter DR, Li R (2005) Variable selection using MM algorithms. *Annals Stat* 33:1617–1642

- [31] Jamshidian M, Jennrich RI (1997) Quasi-Newton acceleration of the EM algorithm. *J Roy Stat Soc B* 59:569–587
- [32] Kent JT, Tyler DE, Vardi Y (1994) A curious likelihood identity for the multivariate t-distribution. *Comm Stat Simulation* 23:441–453
- [33] Kiers HAL (2002) Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Comp Stat Data Anal* 41:157–170
- [34] Kiers HAL, Ten Berge JMF (1992) Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika* 57:371–382
- [35] Lange K (1994) An adaptive barrier method for convex programming. *Methods Applications Analysis* 1:392–402
- [36] Lange K (1995a) A gradient algorithm locally equivalent to the EM algorithm. *J Roy Stat Soc B* 57:425–437
- [37] Lange K (1995b) A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* 5:1–18
- [38] Lange, K (2004) *Optimization*. Springer-Verlag, New York
- [39] Lange K, Fessler JA (1995) Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Trans Image Processing* 4:1430–1438
- [40] Lange K, Hunter DR, Yang I (2000) Optimization transfer using surrogate objective functions (with discussion). *J Comput Graphical Stat* 9:1–20

- [41] Lange K, Little RJA, Taylor JMG (1989) Robust statistical modeling using the t distribution. *JASA* 84:881–896
- [42] Lange K, Wu T (2008) An MM algorithm for multcategory vertex discriminant analysis. *J Computational Graphical Stat* (in press)
- [43] Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791
- [44] Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 13:556–562
- [45] Liao W-H, Huang S-C, Lange K, Bergsneider M (2002) Use of MM algorithm for regularization of parametric images in dynamic PET. *Brain Imaging Using PET*. edited by Senda M, Kimura Yuichi, Herscovitch P, Kimura Yui, Academic Press, New York, pp 107–114
- [46] Little RJA, Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, New York
- [47] Marshall AW, Olkin I (1979), *Inequalities: Theory of Majorization and its Applications*, Academic Press, San Diego
- [48] McLachlan GJ, Krishnan T (1997) *The EM Algorithm and Extensions*. Wiley, New York
- [49] Meng X-L, Rubin DB(1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *J Amer Stat Assoc* 86: 899–909
- [50] Meng X-L, van Dyk D (1997) The EM algorithm – an old folk-song sung to a fast new tune. *J Roy Stat Soc B* 59:511-567

- [51] Meilijson I (1989). A fast improvement to the EM algorithm on its own terms. *J Roy Stat Soc B* 51:127–138
- [52] Ortega JM (1990) *Numerical Analysis: A Second Course*. SIAM, Philadelphia
- [53] Ortega JM, Rheinboldt WC (1970) *Iterative Solutions of Nonlinear Equations in Several Variables*. Academic Press, New York, pp 253–255
- [54] Pauca VP, Piper J, Plemmons RJ (2006) Nonnegative matrix factorization for spectral data analysis. *Linear Algebra Applications* 416:29–47
- [55] Rao CR (1973) *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York
- [56] Sabatti C, Lange K (2002) Genomewide motif identification using a dictionary model. *Proceedings IEEE* 90:1803–1810
- [57] Schölkopf B, Smola AJ (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA
- [58] Steele JM (2004) *The Cauchy-Schwarz Master Class: An Introduction to the Art of Inequalities* Cambridge University Press and the Mathematical Association of America, Cambridge
- [59] Takane Y, Young FW, de Leeuw J (1977) Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika* 42:7-67
- [60] Vaida F (2005) Parameter convergence for EM and MM algorithms. *Stat Sinica* 15:831–840

- [61] Vapnik V (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York
- [62] Varadhan R, Roland C (2008) Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand J Stat* (in press)
- [63] Wu CFJ (1983). On the convergence properties of the EM algorithm. *Annals Stat* 11:95–103