## Title

A Legacy of EM Algorithms



## Journal

International Statistical Review, 90(Suppl 1)

## ISSN

0306-7734

## Authors

Lange, Kenneth

Zhou, Hua

## Publication Date

2022-12-01

## DOI

10.1111/insr.12526

Peer reviewed

HHS Public Access

Author manuscript

*Int Stat Rev*. Author manuscript; available in PMC 2023 May 17.

Published in final edited form as:

*Int Stat Rev*. 2022 December ; 90(Suppl 1): S52–S66. doi:10.1111/insr.12526.

# A Legacy of EM Algorithms

**Kenneth Lange**[1], **Hua Zhou**[2]

[1]Departments of Computational Medicine, Human Genetics, and Statistics, University of California Los Angeles, Los Angeles, 90095-1766CA, USA

[2]Departments of Biostatistics and Computational Medicine, University of California Los Angeles, Los Angeles, 90095-1772CA, USA

## Summary

Nan Laird has an enormous and growing impact on computational statistics. Her paper with Dempster and Rubin on the expectation-maximisation (EM) algorithm is the second most cited paper in statistics. Her papers and book on longitudinal modelling are nearly as impressive. In this brief survey, we revisit the derivation of some of her most useful algorithms from the perspective of the minorisation-maximisation (MM) principle. The MM principle generalises the EM principle and frees it from the shackles of missing data and conditional expectations. Instead, the focus shifts to the construction of surrogate functions via standard mathematical inequalities. The MM principle can deliver a classical EM algorithm with less fuss or an entirely new algorithm with a faster rate of convergence. In any case, the MM principle enriches our understanding of the EM principle and suggests new algorithms of considerable potential in high-dimensional settings where standard algorithms such as Newton's method and Fisher scoring falter.

### Keywords

EM algorithm; MM algorithm; variance component model; longitudinal data analysis

## 1 INTRODUCTION

As of November 2021, the landmark paper on expectation-maximisation (EM) algorithms of Dempster *et al*. (1977) is the second most cited paper across all of statistics, boasting a cumulative count of 64,769 citations according to Google Scholar. This exposition explores variations of the algorithms derived by Dempster *et al*. (1977). These algorithms exemplify some of the most fundamental ideas Nan Laird has contributed to or inspired in statistical science: EM algorithms, the closely related minorisation-maximisation (MM) algorithms, and longitudinal data analysis by mixed models (Garrett *et al*., 2004).

The MM and EM algorithms replace the objective function by a simpler surrogate function. By design, optimising the surrogate function sends the objective function downhill in minimisation and uphill in maximisation. In constructing the surrogate function for an EM algorithm, statisticians rely on notions of missing data. The more general MM algorithm

huazhou@ucla.edu .

calls on skills in inequalities and convex analysis. More often than not, concrete problems also involve parameter constraints. Modern penalty methods incorporate the constraints by imposing penalties on the objective function. A tuning parameter scales the strength of the penalties. In the classical penalty method, the constrained solution is recovered as the tuning parameter tends to infinity. In the augmented Lagrangian method, the constrained solution emerges for a finite value of the tuning parameter.

In the remaining sections, we adopt several notational conventions. Vectors and matrices appear in boldface type; for the most part parameters appear as Greek letters. The differential $df(\boldsymbol{\theta})$ of a scalar-valued function $f(\boldsymbol{\theta})$ equals its row vector of partial derivatives; the transpose $\nabla f(\boldsymbol{\theta})$ of the differential is the gradient. The second differential $d^2 f(\boldsymbol{\theta})$ is the Hessian matrix of second partial derivatives. The Euclidean norm of a vector $\boldsymbol{b}$ and the spectral norm of a matrix $\boldsymbol{A}$ are denoted by $\|\boldsymbol{b}\|$ and $\|\boldsymbol{A}\|$, respectively. All other norms will be appropriately subscripted. The $n$th entry $b_n$ of a vector $\boldsymbol{b}$ must be distinguished from the $n$th vector $\boldsymbol{b}_n$ in a sequence of vectors. To maintain consistency, $b_{ni}$ denotes the $i$th entry of $\boldsymbol{b}_n$. A similar convention holds for sequences of matrices. For symmetric matrices, the relation $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

## 2 THE EM AND MM ALGORITHMS

The numerical analysts Ortega & Rheinboldt (1970) first articulated the MM principle; de Leeuw (1976) saw its potential and created the first MM algorithm of value in statistics. Building on earlier work of Weiszfeld (1937), Voß & Eckhardt (1980) illuminated some convergence properties of MM algorithms. The MM principle currently enjoys its greatest vogue in computational statistics (Hunter & Lange, 2004; Lange *et al.*, 2000; Lange, 2016). The basic idea is to convert a hard optimisation problem into a sequence of simpler ones. In minimisation, the MM principle majorises the objective function $f(\boldsymbol{\theta})$ by a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ anchored at the current point $\boldsymbol{\theta}_n$. Majorisation combines the tangency condition $g(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) = f(\boldsymbol{\theta}_n)$ and the domination condition $g(\boldsymbol{\theta}|\boldsymbol{\theta}_n) \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. The next iterate of the MM algorithm is defined to minimise $g(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$. Because

$$f(\boldsymbol{\theta}_{n+1}) \leq g(\boldsymbol{\theta}_{n+1} \mid \boldsymbol{\theta}_n) \leq g(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_n) = f(\boldsymbol{\theta}_n),$$

the MM iterates generate a deiscent algorithm drivng the objective function downhill. Strictly speaking, the descent property depends only on decreasing $g(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$, not on minimising it. Constraint satisfaction is automatically enforced in finding $\boldsymbol{\theta}_{n+1}$. Under appropriate regularity conditions, an MM algorithm is guaranteed to converge to a local minimum of the objective function (Lange, 2010; Lange *et al.*, 2021). In maximisation, we first minorise and then maximise. Thus, the acronym MM does double duty in the forms majorise-minimise and minorise-maximise.

When it is successful, the MM algorithm simplifies optimisation by (a) separating the variables of a problem, (b) avoiding large matrix inversions, (c) linearising a problem, (d) restoring symmetry, (e) dealing with equality and inequality constraints gracefully, and (f) turning a nondifferentiable problem into a smooth problem. The art in devising an MM

algorithm lies in choosing a tractable surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ that hugs the objective function $f(\boldsymbol{\theta})$ as tightly possible.

The majorisation relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. These rules allow one to work piecemeal in simplifying complicated objective functions. Skill in dealing with inequalities is crucial in constructing majorisations. Classical inequalities such as Jensen's inequality, the information inequality, the arithmetic-geometric mean inequality, and the Cauchy–Schwartz inequality prove useful in many problems. The supporting hyperplane property of a convex function and the quadratic upper bound principle (Böhning & Lindsay, 1988; de Leeuw & Lange, 2009) also find wide application.

The derivation of the EM principle hinges upon a missing data structure. Let $f(\boldsymbol{\theta})$ be the log-likelihood of the observed data with parameter vector $\boldsymbol{\theta}$. In the E step, a surrogate function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ is calculated as the conditional expectation of the complete data log-likelihood given the observed data and the current parameter iterate $\boldsymbol{\theta}_n$. Well-known calculations (Dempster *et al.*, 1977) based on the information inequality demonstrate that the $Q$ function satisfies the domination inequality

$$f(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) - Q(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_n) + f(\boldsymbol{\theta}_n)$$

for all $\boldsymbol{\theta}$. The tangency condition obviously holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_n$. This effectively validates $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ as a minorisation of $f(\boldsymbol{\theta})$ up to an additive constant. Figure 1 displays the $Q$ function of the EM algorithm and the minorisation function of an MM algorithm for the variance component model studied in Section 4. In this example, MM differs from EM, and the MM minorisation function hugs the log-likelihood function tighter than the Q function of the EM algorithm, resulting in faster convergence of MM than EM as depicted later in Figure 2.

## 3 MM ALGORITHMS FOR TRADITIONAL PROBLEMS

Convexity and concavity figure prominently in the construction of many MM algorithms. The supporting hyperplane minorisation

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_n) + df(\boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{x}_n)$$

of a convex function $f(\boldsymbol{x})$ is natural in many applications. For the choice $f(x) = -\ln(x)$, this reads

$$-\ln x \geq -\ln x_n - \frac{x - x_n}{x_n}.$$

Jensen's inequality is instrumental in majorising composite functions of the form $f[u(\boldsymbol{x}) + v(\boldsymbol{x})]$, where $f(y)$ is convex and $u$ and $v$ are positive functions of some underlying parameter vector $\boldsymbol{x}$. In practice, it is often convenient to split the contributions of $u$ and $v$. The majorisation (De Pierro, 1993)

$$f(u+v) \le \frac{u_n}{u_n+v_n} f\left(\frac{u_n+v_n}{u_n}u\right) + \frac{v_n}{u_n+v_n} f\left(\frac{u_n+v_n}{v_n}v\right) \quad (1)$$

achieves this goal. Equality clearly holds whenever $(u, v) = (u_n, v_n)$. For the special case $f(x) = -\ln x$, the minorisation

$$\ln(u+v) \ge \frac{u_n}{u_n+v_n}\ln u + \frac{v_n}{u_n+v_n}\ln v + c_n$$

relies on a constant $c_n$ depending only on $(u_n, v_n)$. This minorisation is handy in splitting log-likelihoods in maximum likelihood estimation with mixture models. It extends to from two to multiple summands. Armed with these ideas, we now explore four examples.

**Example** *Power Series Distributions*

A random variable $X$ concentrated on an interval $[r, \infty)$ of nonnegative integers is said to have a power series distribution if $\Pr(X = k) = \frac{c_k\theta^k}{q(\theta)}$ for all $k \in [r, \infty)$. In this definition, $\theta$ is a positive parameter, the coefficients $c_k$ are nonnegative, and $q(\theta) = \sum_{k=r}^{\infty} c_k\theta^k$ is the appropriate normalising constant (Rao, 1973). Examples include the binomial, negative binomial, Poisson, and logarithmic families and versions of these families truncated at any nonnegative number $r$, especially $r = 1$. For example, the negative binomial distribution has $c_k = \binom{j+k-1}{k}$ and $q(\theta) = (1-\theta)^{-j}$ for $r = 0$ under no truncation. If $x_1, \ldots, x_m$ is a random sample from the power series density and $q(\theta)$ is log-concave, then the log-likelihood of the data is minorised, via the supporting hyperplane inequality, by

$$\begin{aligned} L(\theta) &\ge \sum_{i=1}^{m} x_i \ln\theta - m\ln q(\theta_n) - m[\ln q(\theta_n)]'(\theta-\theta_n) + c_n \\ &= \sum_{i=1}^{m} x_i \ln\theta - m\ln q(\theta_n) - m\frac{q'(\theta_n)}{q(\theta_n)}(\theta-\theta_n) + c_n, \end{aligned}$$

where $c_n$ is a constant independent of $\theta$. Setting the derivative of this surrogate function equal to 0 leads to the MM update

$$\theta_{n+1} = \frac{\bar{x}q(\theta_n)}{q'(\theta_n)},$$

where $\bar{x}$ is the sample average of the observations $x_i$. Anderson *et al.* (2007) derive a straightforward test for log-concavity of $q(\theta)$. Namely, if the coefficients $c_k$ are positive and the ratio $(k+1)c_{k+1}/c_k$ is decreasing in $k$, then $q(\theta)$ is log-concave. The negative binomial fails this test, but the Poisson and binomial distributions qualify. These ideas are pursued in more depth by Wu & Lange (2010).□

**Example** *Cauchy Location and Scale*

The Cauchy density with location $\mu$ and scale $\sigma$ can be written as

$$f(x) = \frac{1}{\pi\sigma\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]}.$$

The usual approach to maximum likelihood estimation of $\mu$ and $\sigma$ involves finding the roots of polynomials of degree $2m - 1$ and $2m$, respectively, for $m$ sample points $x_1, \ldots, x_m$. However, this process tends to be complicated by the existence of multiple local maxima. From the MM perspective, one can exploit the convexity of the function $h(y) = -\log(1 + y)$ via the supporting hyperplane minorisation. If we substitute $\left(\frac{x-\mu}{\sigma}\right)^2$ for $y$, then the log-likelihood is minorised by

$$L(\mu, \sigma) \geq -m \log \sigma - \sum_{i=1}^{m} w_{ni}\left(\frac{x_i - \mu}{\sigma}\right)^2 + c_n$$

$$w_{ni} = \frac{1}{1 + \left(\frac{x_i - \mu_n}{\sigma_n}\right)^2}$$

at iteration $n$, where $c_n$ is an irrelevant constant. The MM algorithm for estimating $\mu$ and $\sigma$ now reduces to weighted least squares with updates

$$\mu_{n+1} = \frac{\sum_{i=1}^{m} w_{ni} x_i}{\sum_{i=1}^{m} w_{ni}} \text{ and } \sigma_{n+1} = \sqrt{\frac{2\sum_{i=1}^{m} w_{ni}(x_i - \mu_{n+1})^2}{m}}.$$

These updates stably increase the likelihood at each iteration. The median of the $x_i$ serves as a starting value for $\mu$. As recommended by Wikipedia, half the sample inter-quartile range is a reasonable starting value for $\sigma$. This example is a special case of the broader algorithm discussed in the next example. □

**Example** *Elliptically Symmetric Distributions*

An elliptically symmetric probability density takes the form

$$f(\boldsymbol{y}) = \frac{e^{-\frac{1}{2}\kappa\left(\delta^2\right)}}{(2\pi)^{\frac{p}{2}}(\det \boldsymbol{\Omega})^{\frac{1}{2}}},$$

where $\boldsymbol{y} \in \mathbb{R}^p$ and $\delta^2 = (\boldsymbol{y} - \boldsymbol{\mu})^* \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$ denotes the Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$. Here, we assume that the function $\kappa(s)$ is strictly increasing and strictly concave and that the matrix $\boldsymbol{\Omega}$ is positive definite. Such densities serve as substitutes for the

multivariate normal distribution in robust estimation (Huber, 2004; Lange *et al.*, 1989; Lange & Sinsheimer, 1993).

Dutter and Huber (Huber, 2004) introduced an MM algorithm driven by the affine majorisation

$$\kappa(t) \leq \kappa(t_n) + \kappa'(t_n)(t - t_n).$$

If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ is a random sample from the density (3), then the multivariate normal log-likelihood

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) = -\frac{1}{2} \sum_{i=1}^{m} \left[ w_{ni} \delta_i^2(\boldsymbol{\theta}) + \ln \det \boldsymbol{\Omega} \right] + c_n$$

with weights $w_{ni} = \kappa' \left[ \delta_i^2(\boldsymbol{\theta}_n) \right]$ and irrelevant constant $c_n$ minorises the log-likelihood of the data under the elliptically symmetric density. The array of techniques from linear algebra for estimating the parameters of a multivariate normal distribution can be brought to bear on maximising $g(\boldsymbol{\theta} | \boldsymbol{\theta}_n)$. For normal/independent distributional families such as the multivariate *t*, the Dutter–Huber algorithm reduces to an EM algorithm (Dempster, 1980; Lange & Sinsheimer, 1993). Given an unstructured mean vector and covariance matrix, the MM updates (Lange *et al.*, 1989; Lange & Sinsheimer, 1993; Little & Rubin, 2019) are

$$\boldsymbol{\mu}_{n+1} = \frac{1}{s_n} \sum_{i=1}^{m} w_{ni} \boldsymbol{y}_i$$

$$\boldsymbol{\Omega}_{n+1} = \frac{1}{m} \sum_{i=1}^{m} w_{ni} (\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1})(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1})^{\top},$$

where $s_n = \sum_{i=1}^{m} w_{ni}$ is the sum of the case weights. For the multivariate *t*, Kent *et al.* (1994) suggest a faster algorithm that replaces the update of $\boldsymbol{\Omega}$ by

$$\boldsymbol{\Omega}_{n+1} = \frac{1}{s_n} \sum_{i=1}^{m} w_{ni} (\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1})(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1})^{\top}.$$

Meng & Van Dyk (1997) justify this amendment within the EM framework.

**Example** *EM for Mixture Models*

Many naive data scientists conflate the EM principle with the EM algorithm for normal mixtures. This level of ignorance is testimony to the importance of this special case. Dempster *et al.* (1977) review the history of the EM clustering algorithm and demonstrate that it possesses the critical ascent property. The EM algorithm makes soft cluster assignments in contrast to the hard assignments of *k*-means clustering. The alternative of

soft choices is possible with admixture models (McLachlan & Krishnan, 2007; Mengersen *et al.*, 2011). An admixture probability density $h(\boldsymbol{y})$ can be written as a convex combination

$$h(\boldsymbol{y}) = \sum_{j=1}^{k} \pi_j h_j(\boldsymbol{y}),$$

where the $\boldsymbol{\pi}_j$ are nonnegative probabilities that sum to 1 and $h_j(\boldsymbol{y})$ is the probability density of group $j$.

Suppose the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ represent a random sample from the admixture density (3). In practice, we want to estimate the admixture proportions $\boldsymbol{\pi}_j$ and whatever further parameters $\boldsymbol{\theta}$ characterise the densities $h_j(\boldsymbol{y}|\boldsymbol{\theta})$. An EM algorithm is natural in this context with group membership as the missing data (Dempster *et al.*, 1977). The EM updates can be derived by invoking the Jensen minorisation (3) for each observation $\boldsymbol{y}_i$ in the form

$$\ln\left[\sum_{j=1}^{k} \pi_j h_j(\boldsymbol{y}_i \mid \boldsymbol{\theta})\right] \geq \sum_{j=1}^{k} w_{nij}[\ln\pi_j + \ln h_j(\boldsymbol{y}_i \mid \boldsymbol{\theta})] + c_n,$$

where $c_n$ is an irrelevant constant and $w_{nij}$ is the posterior probability that $\boldsymbol{y}_i$ belongs to cluster $j$ given the current admixture vector $\boldsymbol{\pi}_n$ and density vector $\boldsymbol{\theta}_n$. Fortunately, this minorisation separates the $\boldsymbol{\pi}$ parameters from the $\boldsymbol{\theta}$ parameters. The problem of maximising the objective $\sum_{j=1}^{k} d_j \ln \pi_j$ for $d_j = \sum_{i=1}^{m} w_{nij}$ is standard with intuitive solution $\boldsymbol{\pi}_{n+1,j} = d_j/m$. Updating the remaining parameters is possible for elliptically symmetric distributions as discussed in the previous example. This involves a second minorisation, which is often ill advised because it generates slowing converging algorithms. It is viable here if we employ a common scale matrix across the groups and the Kent *et al.* (1994) acceleration for the multivariate $t$. □

## 4 MULTI-RESPONSE VARIANCE COMPONENT MODELS

In this example, we contrast the derivations of an EM algorithm and an MM algorithm for the multi-response variance component model. This model involves an $m \times d$ response matrix $\boldsymbol{Y}$ with mean $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{XB}$ and covariance

$$\boldsymbol{\Omega} = \mathrm{Cov}(\mathrm{vec}\,\boldsymbol{Y}) = \sum_{j=1}^{k} \boldsymbol{\Gamma}_j \otimes \boldsymbol{V}_j.$$

The $p \times d$ coefficient matrix $\boldsymbol{B}$ collects the fixed effects, the $d \times d$ covariance matrices $\boldsymbol{\Gamma}_j$ collect the unknown variance components, and the $m \times m$ covariance matrices $\boldsymbol{V}_j$ collect the known variance components. When the vector $\mathrm{vec}\boldsymbol{Y}$ is normally distributed, $\boldsymbol{Y}$ equals a sum of independent matrix normal distributions (Gupta & Nagar, 1999). We now make this assumption and pursue estimation of $\boldsymbol{B}$ and the $\boldsymbol{\Gamma}_j$, which we collectively denote as $\boldsymbol{\Gamma}$. Under the normality assumption, Roth's Kronecker product identity $\mathrm{vec}(\boldsymbol{CDE}) = (\boldsymbol{E}^\top \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{D})$ yields the log-likelihood

$$L(\boldsymbol{B}, \boldsymbol{\Gamma}) = -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}(\operatorname{vec}\boldsymbol{Y} - \boldsymbol{Z}_{\text{vec}}\boldsymbol{B})^{\top}\boldsymbol{\Omega}^{-1}(\operatorname{vec}\boldsymbol{Y} - \boldsymbol{Z}_{\text{vec}}\boldsymbol{B}), \tag{2}$$

where $\boldsymbol{Z} = \boldsymbol{I}_d \otimes \boldsymbol{X}$.

### 4.1 MM Derivation

Updating $\boldsymbol{B}$ given $\boldsymbol{\Gamma}_n$ is accomplished by solving the general least squares problem met earlier in the univariate case. Updating $\boldsymbol{\Gamma}_j$ given $\boldsymbol{B}_n$ is difficult due to the positive semidefiniteness constraint. Typical solutions involve reparameterization of the covariance matrix (Pinheiro & Bates, 1996). The MM algorithm derived in this section gracefully accommodates this constraint.

Updating $\boldsymbol{\Gamma}$ given $\boldsymbol{B}$ requires two minorisations. The convexity of the function $-\ln\det\boldsymbol{\Omega}$ implies the supporting hyperplane minorisation

$$-\frac{1}{2}\ln\det\Omega \geq -\frac{1}{2}\ln\det\Omega_n - \frac{1}{2}\operatorname{tr}\left[\Omega_n^{-1}(\Omega - \Omega_n)\right]. \tag{3}$$

We must also generalise Jensen's majorisation (1). This is accomplished by noting that the function

$$f(\boldsymbol{X}, \boldsymbol{M}) = \begin{cases} \frac{1}{2}\boldsymbol{v}^*\boldsymbol{X}^*\boldsymbol{M}^{-}\boldsymbol{X}\boldsymbol{v} & \boldsymbol{X}\boldsymbol{v} \in \text{Range}(\boldsymbol{M}) \\ \infty & \boldsymbol{X}\boldsymbol{v} \notin \text{Range}(\boldsymbol{M}) \end{cases}$$

is convex for $\boldsymbol{v}$ fixed, where $\boldsymbol{M}$ is positive semidefinite, $\boldsymbol{X}$ is conformable to $\boldsymbol{M}$, and $\boldsymbol{M}^{-}$ is the pseudo-inverse of $\boldsymbol{M}$ (Lange, 2016). Given this fact and the identities $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{AC})\otimes(\boldsymbol{BD})$, $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$, and $\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}$, we have

$$\begin{aligned} \boldsymbol{\Omega}_n\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}_n &= k\left[\frac{1}{k_j}\sum_{j=1}^{k}\boldsymbol{\Gamma}_{nj}\otimes\boldsymbol{V}_j\right]\left[\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\Gamma}_j\otimes\boldsymbol{V}_j\right]^{-1}\left[\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\Gamma}_{nj}\otimes\boldsymbol{V}_j\right] \\ &\leqslant k\sum_{j=1}^{k}\frac{1}{k}(\boldsymbol{\Gamma}_{nj}\otimes\boldsymbol{V}_j)(\boldsymbol{\Gamma}_j\otimes\boldsymbol{V}_j)^{-1}(\boldsymbol{\Gamma}_{nj}\otimes\boldsymbol{V}_j) \\ &= \sum_{j=1}^{k}\left(\boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}\right)\otimes\boldsymbol{V}_j, \end{aligned}$$

or equivalently

$$\boldsymbol{\Omega}^{-1} \leqslant \boldsymbol{\Omega}_n^{-1}\left[\sum_{j=1}^{k}\left(\boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}\right)\otimes\boldsymbol{V}_j\right]\boldsymbol{\Omega}_n^{-1}. \tag{4}$$

This derivation relies on the invertibility of the matrices $\boldsymbol{V}_j$. One can relax this assumption by substituting $\boldsymbol{V}_{\epsilon,j} = \boldsymbol{V}_j + \epsilon\boldsymbol{I}_m$ for $\boldsymbol{V}_j$ and sending $\epsilon$ to 0.

Up to an irrelevant constant, the majorisations (3) and (4) jointly yield the surrogate

$$\begin{aligned} & g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}_{n1}, \ldots, \boldsymbol{\Gamma}_{nk}) \\ &= -\frac{1}{2} \sum_{j=1}^{k} \left\{ \operatorname{tr}\left[\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Gamma}_j \otimes \boldsymbol{V}_j)\right] + \operatorname{vec}(\boldsymbol{R}_n)^\top \left[(\boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}) \otimes \boldsymbol{V}_j\right] \operatorname{vec}(\boldsymbol{R}_n) \right\}, \end{aligned}$$

where $\boldsymbol{R}_n$ is the $m \times d$ matrix satisfying

$$\operatorname{vec}(\boldsymbol{R}_n) = \boldsymbol{\Omega}_n \operatorname{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_n). \tag{5}$$

The first trace here is linear in $\boldsymbol{\Gamma}_j$ and can be expressed as

$$\begin{aligned} & \operatorname{tr}\left[\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Gamma}_j \otimes \boldsymbol{V}_j)\right] = \operatorname{tr}(\boldsymbol{M}_{nj}\boldsymbol{\Gamma}_j) \\ & \boldsymbol{M}_{nj} = (\boldsymbol{I}_d \otimes \boldsymbol{1}_m)^\top \left[(\boldsymbol{1}_d\boldsymbol{1}_d^\top \otimes \boldsymbol{V}_j) \odot \boldsymbol{\Omega}_n^{-1}\right](\boldsymbol{I}_d \otimes \boldsymbol{1}_m), \end{aligned} \tag{6}$$

where $\odot$ takes the Hadamard (pointwise) product of two matrices. To prove this fact, note that if $(\boldsymbol{\Omega}_n^{-1})_{rs}$ is the $(r, s)$th $m \times m$ block of $\boldsymbol{\Omega}_n^{-1}$, then the coefficient of the entry $(\boldsymbol{\Gamma}_j)_{rs}$ is equal to

$$\operatorname{tr}\left[(\boldsymbol{\Omega}_n^{-1})_{rs}\boldsymbol{V}_j\right] = \boldsymbol{1}_m^\top \left[\boldsymbol{V}_j \odot (\boldsymbol{\Omega}_n^{-1})_{rs}\right]\boldsymbol{1}_m.$$

Furthermore, $(\boldsymbol{I}_d \otimes \boldsymbol{1}_m)$ is a diagonal block matrix with each diagonal block equal to $\boldsymbol{1}_m$ and $(\boldsymbol{1}_d\boldsymbol{1}_d^\top \otimes \boldsymbol{V}_j)$ is a block matrix with all blocks equal to $\boldsymbol{V}_j$.

The second trace of $g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}_n)$ simplifies owing to the Kronecker identities $\operatorname{vec}(\boldsymbol{CDE}) = (\boldsymbol{E}^\top \otimes \boldsymbol{C})\operatorname{vec}(\boldsymbol{D})$ and $\operatorname{vec}(\boldsymbol{A})^\top \operatorname{vec}(\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. It follows that the surrogate can be rewritten as

$$\begin{aligned} & g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}_{n1}, \ldots, \boldsymbol{\Gamma}_{nk}) \\ &= -\frac{1}{2} \sum_{j=1}^{k} \left\{ \operatorname{tr}\left[\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Gamma}_j \otimes \boldsymbol{V}_j)\right] + \operatorname{tr}\left(\boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}^\top\right) \right\} \\ &= -\frac{1}{2} \sum_{j=1}^{k} \left\{ \operatorname{tr}(\boldsymbol{M}_{nj}\boldsymbol{\Gamma}_j) + \operatorname{tr}\left(\boldsymbol{\Gamma}_{nj}^\top \boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\right) \right\}. \end{aligned} \tag{7}$$

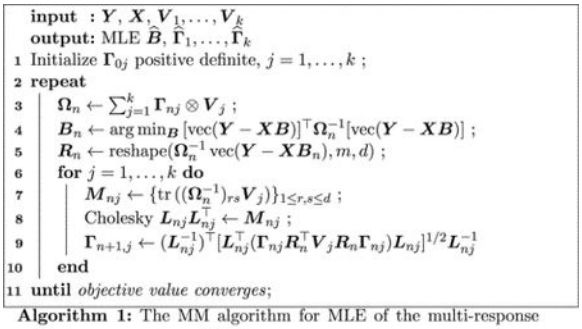
The directional derivative of $g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}_{n1}, \ldots, \boldsymbol{\Gamma}_{nk})$ with respect to $\boldsymbol{\Gamma}_j$ in the direction $\boldsymbol{\Delta}_j$ is

$$\begin{aligned} & -\frac{1}{2}\operatorname{tr}(\boldsymbol{M}_{nj}\boldsymbol{\Delta}_j) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Gamma}_{nj}^\top \boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j\boldsymbol{\Gamma}_j^{-1}\right) \\ &= -\frac{1}{2}\operatorname{tr}(\boldsymbol{M}_{nj}\boldsymbol{\Delta}_j) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}\boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j\right). \end{aligned}$$

Because all directional derivatives of the surrogate vanish at a stationary point, the matrix equation

$$\boldsymbol{M}_{nj} = \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Gamma}_{nj}\boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}\boldsymbol{\Gamma}_j^{-1} \tag{8}$$

holds. Fortunately, this equation admits an explicit solution. For positive scalers $a$ and $b$, the solution to the equation $b = 1x^{-1}ax^{-1}$ is $x = \pm\sqrt{a/b}$ The matrix analogue of this equation is the Riccati equation $\boldsymbol{B} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}$, whose solution is summarised in the next lemma.




**Algorithm 1:** The MM algorithm for MLE of the multi-response variance component model.

**Lemma 1**. *Assume* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are positive definite and* $\boldsymbol{L}$ *is the Cholesky factor of* $\boldsymbol{B}$*. Then* $\boldsymbol{Y} = (\boldsymbol{L}^{-1})^\top (\boldsymbol{L}^\top \boldsymbol{A}\boldsymbol{L})^{1/2}\boldsymbol{L}^{-1}$ *is the unique positive definite solution to the matrix equation* $\boldsymbol{B} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}$.

The Cholesky factor $\boldsymbol{L}$ in Lemma 4.1 can be replaced by the symmetric square root of $\boldsymbol{B}$. The solution, which is unique, remains the same. The Cholesky decomposition is preferred for its cheaper computational cost and better numerical stability.

Algorithm 1 summarises the MM algorithm for fitting the multi-response model (1). Each iteration invokes $k$ Cholesky decompositions and symmetric square roots of $d \times d$ positive definite matrices. Fortunately in most applications, $d$ is a small number.

### 4.2 EM Derivation

The landmark paper (Dempster *et al.*, 1977) by Nan Laird and co-authors features the EM derivation for variance component models with univariate response. Later extensions to multivariate responses include Reinsel (1984) and Glanz & Carvalho (2018). We give a self-contained derivation here for ease of comparison with the MM algorithm. Derivation of the EM algorithm hinges upon the missing data and conditional expectation. If the response matrix $\boldsymbol{Y}$ can be written as the sum $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{Z}_1 + \ldots + \boldsymbol{Z}_k$ of independent random matrices with $\mathrm{vec}\boldsymbol{Z}_j \sim N(\boldsymbol{0}, \boldsymbol{\Omega}_j)$, then $\mathrm{vec}\boldsymbol{Y} \sim N(\mathrm{vec}(\boldsymbol{X}\boldsymbol{\beta}); \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \sum_{j=1}^k \boldsymbol{\Omega}_j$. Under the matrix normal assumption, $\boldsymbol{\Omega}_j = \boldsymbol{\Gamma}_j \otimes \boldsymbol{V}_j$. The complete data log-likelihood for the unobserved $\boldsymbol{Z}_j$ is

$$-\frac{1}{2}\sum_{j=1}^{k} \ln\det{}^{+}\boldsymbol{\Omega}_j - \frac{1}{2}\sum_{j=1}^{k} (\mathrm{vec}\boldsymbol{Z}_j)^T \boldsymbol{\Omega}_j^{+} (\mathrm{vec}\boldsymbol{Z}_j),$$

where $\det^{+}\boldsymbol{\Omega}_j$ denotes the pseudo-determinant of $\boldsymbol{\Omega}_j$ and $\boldsymbol{\Omega}_j^{+}$ the pseudo-inverse of $\boldsymbol{\Omega}_j$. To compute the surrogate function for the EM algorithm, one needs the conditional expectations

$$\boldsymbol{E}_{nj} = \mathrm{E}(\mathrm{vec}\boldsymbol{Z}_j \mid \boldsymbol{Y}, \boldsymbol{\theta}_n) = \boldsymbol{\Omega}_{nj}\boldsymbol{\Omega}_n^{-1}\mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_n)$$

and the conditional covariances

$$\boldsymbol{F}_{nj} = \mathrm{Cov}(\mathrm{vec}\,\boldsymbol{Z}_j \mid \boldsymbol{Y}, \boldsymbol{\theta}_n) = \boldsymbol{\Omega}_{nj} - \boldsymbol{\Omega}_{nj}\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Omega}_{nj},$$

where $\boldsymbol{\theta}$ is the parameter vector. These are employed to compute the conditional second moments

$$\boldsymbol{G}_{nj} = \mathrm{E}\left[(\mathrm{vec}\,\boldsymbol{Z}_j)(\mathrm{vec}\,\boldsymbol{Z}_j)^\top \mid \boldsymbol{Y}, \boldsymbol{\theta}_n\right] = \boldsymbol{F}_{nj} + \boldsymbol{E}_{nj}(\boldsymbol{E}_{nj})^\top.$$

Here, the random vector $\boldsymbol{Z}_j$ should be replaced by $\boldsymbol{Z}_k - \boldsymbol{X}\boldsymbol{B}_n$ when $j = k$.

One can readily check that $\boldsymbol{\Omega}_j^+ = \boldsymbol{\Gamma}_j^+ \otimes \boldsymbol{V}_j^+ = \boldsymbol{\Gamma}_j^{-1} \otimes \boldsymbol{V}_j^+$ for $\boldsymbol{\Gamma}_j$ invertible. Because the pseudo-determinant of a positive semidefinite matrix equals the product of its positive eigenvalues, the formulas

$$\det{}^+ \boldsymbol{\Omega}_j = (\det \boldsymbol{\Gamma}_j)^{r_i}\left(\det{}^+ \boldsymbol{V}_j\right)^{s_i}$$

$$\ln\det{}^+ \boldsymbol{\Omega}_j = r_j \ln \det \boldsymbol{\Gamma}_j + s_i \ln\det{}^+ \boldsymbol{V}_j$$

apply, where $r_j = \mathrm{rank}(\boldsymbol{V}_j^+)$ and $s_j = \mathrm{rank}(\boldsymbol{\Gamma}_j^+)$. In the M step of the EM algorithm, one maximises the surrogate

$$-\frac{1}{2}\sum_{j=1}^{k} r_j \ln \det \boldsymbol{\Gamma}_j - \frac{1}{2}\sum_{j=1}^{k} \mathrm{tr}\left[(\boldsymbol{\Gamma}_j^{-1} \otimes \boldsymbol{V}_j^+)\boldsymbol{G}_{nj}\right]. \tag{9}$$

For $\boldsymbol{\Gamma}_j$ unstructured, we substitute $\boldsymbol{\Lambda}_j = \boldsymbol{\Gamma}_j^{-1}$ and maximise with respect to $\boldsymbol{\Lambda}_j$. Fortunately, the next lemma can be invoked.

**Lemma 2**. *If the matrices* $\boldsymbol{A}$, $\boldsymbol{B}$, *and* $\boldsymbol{C}$ *are* $d \times d$, $m \times m$, *and* $dm \times dm$ *respectively, then*

$$\mathrm{tr}\left[(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{C}^\top\right] = \mathrm{tr}\left\{(\boldsymbol{I}_d \otimes \boldsymbol{1}_m)^\top\left[(\boldsymbol{1}_d\boldsymbol{1}_d^\top \otimes \boldsymbol{B}) \odot \boldsymbol{C}\right](\boldsymbol{I}_d \otimes \boldsymbol{1}_m)\boldsymbol{A}^\top\right\}.$$

**Proof** *This trace identity is essentially proved in our derivation of the corresponding MM algorithm in* Section 4.1. □

Lemma 4.2 yields

$$\mathrm{tr}\left[(\boldsymbol{\Lambda}_j \otimes \boldsymbol{V}_j^+)\boldsymbol{G}_j^{(t)}\right] = \mathrm{tr}\left\{(\boldsymbol{I}_d \otimes \boldsymbol{1}_m)^\top\left[(\boldsymbol{1}_d\boldsymbol{1}_d^\top \otimes \boldsymbol{V}_j^+) \odot \boldsymbol{G}_j^{(t)}\right](\boldsymbol{I}_d \otimes \boldsymbol{1}_m)\boldsymbol{\Lambda}_j\right\}.$$

The stationarity condition

$$\mathbf{0} = \frac{1}{2} r_j \boldsymbol{\Lambda}_j^{-1} - \frac{1}{2} (\boldsymbol{I}_d \otimes \mathbf{1}_m)^\top \left[ \left( \mathbf{1}_d \mathbf{1}_d^\top \otimes \boldsymbol{V}_j^+ \right) \odot \boldsymbol{G}_j^{(i)} \right] (\boldsymbol{I}_d \otimes \mathbf{1}_m)$$

now entails the update

$$\begin{aligned} \boldsymbol{\Gamma}_{n+1,j} &= r_j^{-1} (\boldsymbol{I}_d \otimes \mathbf{1}_m)^\top \left[ \left( \mathbf{1}_d \mathbf{1}_d^\top \otimes \boldsymbol{V}_j^+ \right) \odot \boldsymbol{G}_{nj} \right] (\boldsymbol{I}_d \otimes \mathbf{1}_m) \\ &= \boldsymbol{\Gamma}_{nj} - r_j^{-1} \boldsymbol{\Gamma}_{nj} \boldsymbol{M}_{nj} \boldsymbol{\Gamma}_{nj} + r_j^{-1} \boldsymbol{\Gamma}_{nj} \boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}, \end{aligned} \tag{10}$$

where $\boldsymbol{M}_{nj}$ is the $d \times d$ matrix defined by (6) and $\boldsymbol{R}_n$ is the $m \times d$ matrix defined by (5). The second equation invokes the identities $\boldsymbol{V}_j \boldsymbol{V}_j^+ \boldsymbol{V}_j = \boldsymbol{V}_j$, $\operatorname{tr}(\boldsymbol{V}_j \boldsymbol{V}_j^+) = \operatorname{rank}(\boldsymbol{V}_j)$, and the cyclic permutation property of the trace.

In the case $k = 1$, the single update reduces to

$$\boldsymbol{\Gamma}_{n+1} = \frac{1}{r} (\boldsymbol{Y} - \boldsymbol{X} \boldsymbol{B}_n)^\top \boldsymbol{V}^+ (\boldsymbol{Y} - \boldsymbol{X} \boldsymbol{B}_n),$$

which matches the earlier result of Glanz & Carvalho (2018). When $\boldsymbol{\Gamma}_j$ is the scalar $\sigma_j^2$, the update (10) reduces to the classical update (Dempster *et al.*, 1977)

$$\sigma_{n+1,j}^2 = \sigma_{n,j}^2 - \frac{\sigma_{n,j}^4}{r_j} \left[ \operatorname{tr}\left( \boldsymbol{\Omega}_n^{-1} \boldsymbol{V}_j \right) - (\boldsymbol{y} - \boldsymbol{X} \boldsymbol{\beta}_n)^\top \boldsymbol{\Omega}_n^{-1} \boldsymbol{V}_j \boldsymbol{\Omega}_n^{-1} (\boldsymbol{y} - \boldsymbol{X} \boldsymbol{\beta}_n) \right].$$

Algorithm 2 summarises the EM algorithm for fitting the multi-response model (2). The additive update of $\boldsymbol{\Gamma}_j$ in the EM algorithm differs markedly from the multiplicative update in the MM algorithm. The computational cost of each EM iteration is similar to that of MM. Both are dominated by the inversion of the $md \times md$ covariance matrix $\boldsymbol{\Omega}_n$.

For the univariate response case $d = 1$, Zhou *et al.* (2019) show that the MM algorithm enjoys a faster convergence rate than EM. Here, we verify the same behaviour empirically for a variance component model with $d = 4$ responses for $m = 500$ subjects, $k = 3$ variance components, and $p = 3$ fixed effect covariates. We start both algorithms from the same initial point in each of 100 simulation replicates. Figure 2 shows that MM algorithm converges faster than EM in all replicates.

**input** : $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_k$
**output:** MLE $\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Gamma}}_1, \ldots, \widehat{\boldsymbol{\Gamma}}_k$
Initialize $\boldsymbol{\Gamma}_{0j}$ positive definite, $j = 1, \ldots, k$ ;
**repeat**
  $\boldsymbol{\Omega}_{nj} \leftarrow \boldsymbol{\Gamma}_{nj} \otimes \boldsymbol{V}_j$ ;
  $\boldsymbol{\Omega}_n \leftarrow \sum_{j=1}^k \boldsymbol{\Omega}_{nj}$ ;
  $\boldsymbol{B}_n \leftarrow \arg\min_{\boldsymbol{B}} [\operatorname{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})]^\top \boldsymbol{\Omega}_n^{-1} [\operatorname{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})]$ ;
  $\boldsymbol{R}_n \leftarrow \operatorname{reshape}(\boldsymbol{\Omega}_n^{-1} \operatorname{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_n), m, d)$ ;
  **for** $j = 1, \ldots, k$ **do**
    $\boldsymbol{M}_{nj} \leftarrow \{\operatorname{tr}((\boldsymbol{\Omega}_n^{-1})_{rs} \boldsymbol{V}_j)\}_{1 \le r,s \le d}$ ;
    $\boldsymbol{\Gamma}_{n+1,j} \leftarrow \boldsymbol{\Gamma}_{nj} - r_j^{-1} \boldsymbol{\Gamma}_{nj} \boldsymbol{M}_{nj} \boldsymbol{\Gamma}_{nj} + r_j^{-1} \boldsymbol{\Gamma}_{nj} \boldsymbol{R}_n^\top \boldsymbol{V}_j \boldsymbol{R}_n \boldsymbol{\Gamma}_{nj}$
  **end**
**until** *objective value converges*;

**Algorithm 2:** The EM algorithm for MLE in the multi-response variance component model.

### 4.3 A Problem Involving a Moore–Penrose Inverse

The derivations so far assume that the variance components $\boldsymbol{\Gamma}_j$ are unstructured covariance matrices with $kd(d+1)/2$ parameters. Under a limited sample size, $\boldsymbol{\Gamma}_j$ cannot be estimated reliably, especially when the numbers of responses $d$ and variance components $k$ are large. A more parsimonious model imposes a low rank structure on all $\boldsymbol{\Gamma}_j$ except that associated with $\boldsymbol{V}_k = \boldsymbol{I}_m$. Then maximising the MM surrogate function (7) boils down to the problem of minimizing

$$\operatorname{tr}(\boldsymbol{M}\boldsymbol{X}) + \operatorname{tr}\left(\boldsymbol{N}\boldsymbol{X}^{+}\right)$$

where $\boldsymbol{M}$ and $\boldsymbol{N}$ are known $d \times d$ positive definite matrices and $\boldsymbol{X}$ is a positive semidefinite $d \times d$ matrix of rank $r \leq d$. Denote the thin eigendecomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\top}$, where $\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}_r$ and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_i > 0$. We first determine the optimal eigenvalues $\sigma_i$ and then eigenvectors $\boldsymbol{U}$. The objective is

$$\begin{aligned}
&\operatorname{tr}\left(\boldsymbol{U}^{\top}\boldsymbol{M}\boldsymbol{U}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{U}^{\top}\boldsymbol{N}\boldsymbol{U}\boldsymbol{\Sigma}^{+}\right) \\
&= \operatorname{tr}(\boldsymbol{A}\Sigma) + \operatorname{tr}\left(\boldsymbol{B}\Sigma^{+}\right) \\
&= \sum_{i=1}^{r} \sigma_i a_{ii} + \sum_{i=1}^{r} \sigma_i^{-1} b_{ii},
\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{U}^{\top}\boldsymbol{M}\boldsymbol{U}$, $\boldsymbol{B} = \boldsymbol{U}^{\top}\boldsymbol{N}\boldsymbol{U}$. Setting the derivatives to zero yields the optimal eigenvalues $\sigma_i = \sqrt{b_{ii}/a_{ii}}$.

Now the task is to minimise

$$f(\boldsymbol{U}) = 2\sum_{i=1}^{r}\sqrt{a_{ii}b_{ii}} = 2\sum_{i=1}^{r}\sqrt{\boldsymbol{u}_i^{\top}\boldsymbol{M}\boldsymbol{u}_i\boldsymbol{u}_i^{\top}\boldsymbol{N}\boldsymbol{u}_i}$$

under the orthogonality constraint $\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}_r$, which is subject to a suite of algorithms for manifold optimisation such as Manopt (Boumal *et al.*, 2014) or a simple projected gradient descent algorithm. We record the gradient as

$$\nabla_{\boldsymbol{u}_i} f(\boldsymbol{U}) = 2\sqrt{\frac{\boldsymbol{u}_i^{\top}\boldsymbol{N}\boldsymbol{u}_i}{\boldsymbol{u}_i^{\top}\boldsymbol{M}\boldsymbol{u}_i}}\boldsymbol{M}\boldsymbol{u}_i + 2\sqrt{\frac{\boldsymbol{u}_i^{\top}\boldsymbol{M}\boldsymbol{u}_i}{\boldsymbol{u}_i^{\top}\boldsymbol{N}\boldsymbol{u}_i}}\boldsymbol{N}\boldsymbol{u}_i\,.$$

Alternatively, split variables by replacing $\boldsymbol{U}$ by $\boldsymbol{A}$ and $\boldsymbol{U}^{\top}$ by $\boldsymbol{B}^{\top}$. Then impose the constraints $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{B}^{\top}\boldsymbol{A} = \boldsymbol{I}_r$. The penalised objective

$$\operatorname{tr}\left(\boldsymbol{B}^{\top}\boldsymbol{M}\boldsymbol{A}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{B}^{\top}\boldsymbol{N}\boldsymbol{A}\boldsymbol{\Sigma}^{+}\right) + \frac{\rho}{2}\left\|\boldsymbol{B}^{\top}\boldsymbol{A} - \boldsymbol{I}_r\right\|_F^2 + \frac{\rho}{2}\left\|\boldsymbol{A} - \boldsymbol{B}\right\|_F^2$$

for $\rho > 0$ large has differential with respect to $\boldsymbol{A}$ of

$$\boldsymbol{\Sigma}\boldsymbol{B}^\top\boldsymbol{M} + \boldsymbol{\Sigma}^+\boldsymbol{B}^\top\boldsymbol{N} + \rho\left(\boldsymbol{B}^\top\boldsymbol{A} - \boldsymbol{I}_r\right)^\top\boldsymbol{B}^\top + \rho(\boldsymbol{A} - \boldsymbol{B})^\top.$$

Set this equal to $\mathbf{0}$ and solve for $\boldsymbol{A}$ in the form

$$\boldsymbol{A}^\top = -\left(\rho^{-1}\boldsymbol{\Sigma}\boldsymbol{B}^\top\boldsymbol{M} + \rho^{-1}\boldsymbol{\Sigma}^+\boldsymbol{B}^\top\boldsymbol{N} - 2\boldsymbol{B}^\top\right)\left(\boldsymbol{I}_d + \boldsymbol{B}\boldsymbol{B}^\top\right)^{-1}.$$

A similar update holds for $\boldsymbol{B}$.

## 5 DISCUSSION

The senior (citizen) author of this paper remembers being mesmerised by Nan Laird's EM seminar at UCLA in the late 1970s. Nan opened an entirely new toolbox of optimisation. The beautiful abstraction and generality of the EM principle has served the statistics community well for decades. The principle is capable of generating maximum likelihood algorithms motivated by intermediate quantities of natural statistical interest. It is worth stressing that EM and Fisher scoring are unique contributions by statisticians to optimisation practice. However, there is no panacea in optimisation. Each problem class presents unique challenges and deserves to be attacked from a variety of perspectives. Often hybrid algorithms work best.

The MM principle distils the essence of EM and frees it from the sometimes elusive notion of missing data. As we have witnessed, EM and MM algorithms for the same problem do not necessarily coincide. When they differ, their rates of convergence and computational complexity can also differ. Our exposition of EM and MM algorithms for variance component models illustrates these points. In this case, the MM algorithm appears faster.

The current paper offers, at best, a snapshot of the current state of the MM art. New applications are in the pipeline. Our recent research on constrained optimisation shows how the MM principle, set projection, and the Courant penalty method can cooperate to solve constrained problems involving nonconvexity and sparsity (Chi *et al.*, 2014; Keys *et al.*, 2019; Landeros *et al.*, 2022; Xu *et al.*, 2017). Many challenges remain in theory, numerical practice, and software development. Fortunately, current researchers stand on the shoulders of giants such as Nan Laird and Jan de Leeuw in attacking these issues. We are profoundly grateful to Nan for her many advances in computational statistics. Only a handful of statisticians can claim a legacy of such distinction.

## ACKNOWLEDGEMENTS

This research was supported in part by NIH grants HG006139 and GM14179 and by NSF grants DMS-2054253 and IIS-2205441.

## REFERENCES


Anderson GD, Vamanamurthy MK & Vuorinen M (2007). Generalized convexity and inequalities. J. Math. Anal. Appl, 335(2), 1294–1308.

Böhning D & Lindsay BG (1988). Monotonicity of quadratic-approximation algorithms. Ann. Inst. Statist. Math, 40(4), 641–663. 10.1007/BF00049423

Boumal N, Mishra B, Absil P-A & Sepulchre R (2014). Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res, 15(42), 1455–1459. https://www.manopt.org

Chi EC, Zhou H & Lange K (2014). Distance majorization and its applications. Math. Program, 146(1–2), 409–436. [PubMed: 25392563]

de Leeuw J (1976). Applications of convex analysis to multidimensional scaling Elsevier.

de Leeuw J & Lange K (2009). Sharp quadratic majorization in one dimension. Comput. Stat. Data Anal, 53(7), 2471–2484. [PubMed: 21738282]

De Pierro AR (1993). On the relation between the isra and the em algorithm for positron emission tomography. IEEE Trans. Med. Imaging, 12(2), 328–333. [PubMed: 18218422]

Dempster AP (1980). Iterative reweighted least squares for linear regression when errors are normal/ inde-pendent distributed. Multivariate Anal., 1980, 35–57.

Dempster AP, Laird NM & Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B, 39(1–38).

Garrett MF, Laird NM & Ware JH (2004). Applied longitudinal analysis. Wiley-Interscience: Hoboken (NJ).

Glanz H & Carvalho L (2018). An Expectation–Maximization algorithm for the matrix normal distribution with an application in remote sensing. J. Multivariate Anal, 167, 31–48.

Gupta AK & Nagar DK (1999). Matrix variate distributions, Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.

Huber PJ (2004). Robust statistics, Vol. 523. John Wiley & Sons.

Hunter DR & Lange K (2004). A tutorial on MM algorithms. Amer. Statist, 58(1), 30–37. 10.1198/0003130042836

Kent JT, Tyler DE & Vard Y (1994). A curious likelihood identity for the multivariate t-distribution. Commun. Stat.-Simul. Comput, 23(2), 441–453.

Keys KL, Zhou H & Lange K (2019). Proximal distance algorithms: theory and practice. J. Mach. Learn. Res, 20(66), 1–38.

Landeros A, Padilla OM, Zhou H & Lange K (2022). Extensions to the proximal distance method of constrained optimization. J. Mach. Learn. Res, 23(182), 1–45.

Lange K (2010). Numerical Analysis for Statisticians, Second, Statistics and Computing. Springer: New York. 10.1007/978-1-4419-5945-4

Lange K (2016). MM optimization algorithms. Society for Industrial and Applied Mathematics: Philadelphia, PA. 10.1137/1.9781611974409

Lange K, Hunter DR & Yang I (2000). Optimization transfer using surrogate objective functions. J. Comput. Graph. Statist, 9(1), 1–59. With discussion, and a rejoinder by Hunter and Lange.

Lange KL, Little RJA & Taylor J.eremyM.G. (1989). Robust statistical modeling using the t distribution. J. Am. Stat. Assoc, 84(408), 881–896.

Lange K & Sinsheimer JS (1993). Normal/independent distributions and their applications in robust regression. J. Comput. Graph. Stat, 2, 175–198.

Lange K, Won J-H, Landeros A & Zhou H 2021. Nonconvex optimization via MM algorithms: Convergence theory. In Computational statistics in data science, American Cancer Society, pp. 1–22. 10.1002/9781118445112.stat08295

Little RJA & Rubin DB (2019). Statistical analysis with missing data, Vol. 793. John Wiley & Sons.

McLachlan GJ & Krishnan T (2007). The EM algorithm and extensions, Vol. 382. John Wiley & Sons.

Meng X-L & Van Dyk D (1997). The EM algorithm an old folk-song sung to a fast new tune. J. R. Stat. Soc.: Ser. B (Stat. Methodol.), 59(3), 511–567.

Mengersen KL, Robert C & Titterington M (2011). Mixtures: Estimation and applications, Vol. 896. John Wiley & Sons.

Ortega JM & Rheinboldt WC (1970). Iterative solution of nonlinear equations in several variables. Academic Press: New York-London.

Pinheiro J & Bates D (1996). Unconstrained parametrizations for variance-covariance matrices. Stat. Comput, 6(3), 289–296 English. 10.1007/BF00140873

Rao CR (1973). Linear Statistical Inference and its Applications, 2nd ed. John Wiley & Sons.

Reinsel G (1984). Estimation and prediction in a multivariate random effects generalized linear model. J. Amer. Statist. Assoc, 79(386), 406–414.

Voß H & Eckhardt U (1980). Linear convergence of generalized Weiszfeld's method. Computing, 25(3), 243–251.

Weiszfeld E (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Math. J., First Ser, 43, 355–386.

Wu TT & Lange K (2010). The MM alternative to EM. Stat. Sci, 25(4), 492–505.

Xu J, Chi E & Lange K (2017). Generalized linear model regression under distance-to-set penalties. In Advances in neural information processing systems, pp. 1385–1395.

Zhou H, Hu L, Zhou J & Lange K (2019). MM algorithms for variance components models. J. Comput. Graph. Statist, 28(2), 350–361.

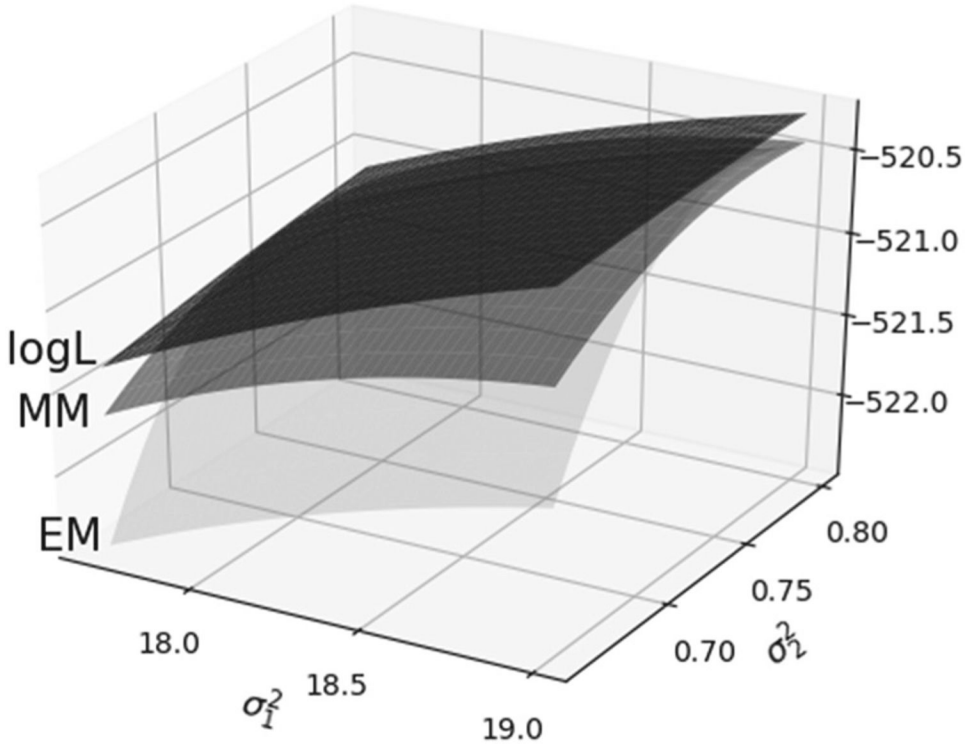



**Figure 1.**

The Q function of EM and the corresponding MM surrogate minorise the log-likelihood surface of a univariate response, two variance component model at the point (18.5, 0.7). In this example, the MM surrogate hugs the log-likelihood surface tighter than the EM surrogate

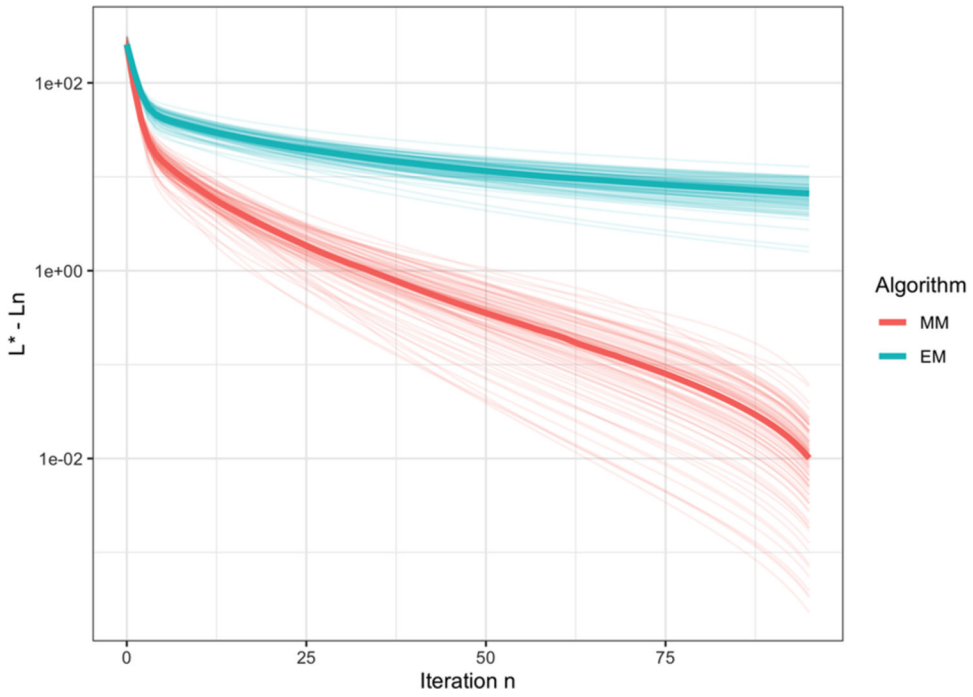



**Figure 2.**

MM algorithm converges faster than EM for a multi-response variance component model with d = 4 responses for m = 500 subjects, k = 3 variance components, and p = 3 fixed effect covariates. $L^{\star} - L_{\mathrm{n}}$ indicates the difference in log-likelihood between the found MLE and the nth iterate. EM and MM algorithms start from the same point in each of the 100 simulation replicates