Notes on the MLE in curved exponential families

by ML Eaton and DA Freedman December 18/00

For $i = 1, 2$ and $j = 1, \ldots, n$, let Y_{ij} be independent normal random variables, with common expectation α. For $i = 1, 2$, the variance is $\sigma_i^2 > 0$. We refer to the Y_{ij} as "observations," and the full collection $\{Y_{ij} : i = 1, 2,\ j = 1, \ldots, n\}$ is a "sample"; the sub-collections with $i = 1$ or 2 are "sub-samples." As a matter of notation, α and σ_i^2 are arguments in the likelihood function, as well as the true values from which the sample is drawn. When the distinction matters, the true values will be denoted by underscores, as $\underline{\alpha}$, $\underline{\sigma_i^2}$. Let

$$y_i = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - y_i)^2, \tag{1}$$

the mean and variance of the sub-samples with $i = 1, 2$; we use the old-fashioned denominator n for the variances. We consider only samples with

$$n > 1,\ y_1 \neq y_2,\ s_1^2 > 0, \text{ and } s_2^2 > 0. \tag{2}$$

Proposition 1. For some samples satisfying (2), the log likelihood function has two maxima. On the other hand, for given $\underline{\alpha}$ and $\underline{\sigma_i^2}$, there is an exceptional set of samples whose probability decreases at a geometric rate as $n \to \infty$; for all other samples, the log likelihood function is unimodal, and the feasible-GLS algorithm converges to this maximum.

Argument. Up to an additive constant, the log likelihood function is

$$\begin{aligned} L(\alpha, \sigma_1, \sigma_2) &= -\sum_{j=1}^{n} \frac{(y_{1j} - \alpha)^2}{2\sigma_1^2} - \sum_{j=1}^{n} \frac{(y_{2j} - \alpha)^2}{2\sigma_2^2} - n\log\sigma_1 - n\log\sigma_2 \\ &= -\frac{n}{2}\Big\{ \frac{s_1^2 + (y_1 - \alpha)^2}{\sigma_1^2} + \frac{s_2^2 + (y_2 - \alpha)^2}{\sigma_2^2} + \log\sigma_1^2 + \log\sigma_2^2 \Big\}, \end{aligned} \tag{3}$$

where y_i and s_i^2 are defined by (1).

(4) Fix α. The log likelihood function is strictly concave in $(1/\sigma_1^2, 1/\sigma_2^2)$. The maximum in σ_i is attained when $\sigma_i^2 = s_i^2 + (y_i - \alpha)^2$.

Substitute these $\sigma_i(\alpha)$ into (3):

$$M(\alpha) = -\frac{n}{2}\Big\{ \log\left[s_1^2 + (y_1 - \alpha)^2\right] + \log\left[s_2^2 + (y_2 - \alpha)^2\right] \Big\}, \tag{5}$$

up to an additive constant. Hence,

$$M(\pm\infty) = -\infty. \tag{6}$$

Differentiate (5) with respect to α:

$$M'(\alpha) = \frac{n(y_1 - \alpha)}{s_1^2 + (y_1 - \alpha)^2} + \frac{n(y_2 - \alpha)}{s_2^2 + (y_2 - \alpha)^2}. \tag{7}$$

By change of scale, including perhaps change of sign, we can take $y_1 = 0$ and $y_2 = 1$. In the new scale, the sample variance of the first group of data is $a^2 > 0$, the second variance is $b^2 > 0$, and the common expectation becomes μ. The scale is data-dependent. Now rewrite $M'(\alpha)$ in (7) as

$$n \frac{1}{a^2 + \mu^2} \frac{1}{b^2 + (1 - \mu)^2} Q(\mu) \quad \text{where} \quad Q(\mu) = -2\mu^3 + 3\mu^2 - (a^2 + b^2 + 1)\mu + a^2. \tag{8}$$

Of course, $Q(-\infty) = \infty$ and $Q(\infty) = -\infty$. Moreover, Q has either one real root or three.

We claim that for sufficiently small a^2 and b^2, e.g., $a^2 \le 0.01$ and $b^2 \le 0.01$, there are three real roots. Indeed, the limiting polynomial

$$Q_0(\mu) = -2\mu^3 + 3\mu^2 - \mu = -\mu(\mu - 1)(2\mu - 1)$$

has roots 0, 1/2, 1; between 0 and 1/2, this polynomial is negative, with a minimum of about -0.1; between 1/2 and 1, the polynomial is positive, with a maximum of about 0.1. Moreover,

$$Q(\mu) \ge 6 \text{ for } \mu \le -1 \quad \text{and} \quad Q(\mu) \le -6 + a^2 \text{ for } \mu \ge 2. \tag{9}$$

Hence, if $a^2 > 0$ and $b^2 > 0$ are small, Q will have three real roots; necessarily, Q changes sign at each root. The corresponding changes of sign in M' imply that M has two local maxima, separated by a local minimum. There are two local maxima in the log likelihood function L, and a third critical point which corresponds to a saddle point in L.

In the other direction, $Q'(\mu) = -6\mu^2 + 6\mu - (1 + a^2 + b^2)$, which is negative for all μ provided

$$6^2 - 4 \times 6 \times (1 + a^2 + b^2) < 0,$$

i.e., $12 - 24(a^2 + b^2) < 0$, or $a^2 + b^2 > 1/2$. In original units, this means

$$s_1^2 + s_2^2 > \frac{1}{2}|y_1 - y_2|. \tag{10}$$

If so, Q in (8) has only one real root, i.e., the likelihood equation has only one real root, and the likelihood is unimodal, which proves

(11) If (10) holds, the likelihood equation has only one real root.

Clearly, the chance that (10) obtains is bounded below by $1 - A\rho^n$ where $0 < A < \infty$ and $0 < \rho < 1$ depend on the true $\underline{\alpha}$, $\underline{\sigma}_i^2$ but not on n, the number of observations in each subsample; nor do A and ρ depend on the arguments of the likelihood function—otherwise, insanity prevails.

For feasible GLS, if σ_1 and σ_2 are fixed, the log likelihood function (3) is strictly concave in α, with its maximum at

$$\alpha = \frac{\sigma_1^{-2} y_1 + \sigma_2^{-2} y_2}{\sigma_1^{-2} + \sigma_2^{-2}}. \tag{12}$$

On the other hand, if α is fixed, maximization in σ_i is governed by (4). We iterate by alternating these maximizations, starting for instance with $\alpha = \alpha_0$. The σ_i computed from α_0 via (4) will be denoted $\sigma_{i,0}$, and so forth. The log likelihood increases monotonically; let L_∞ be the limit.

Suppose by way of contradiction that $L_\infty < \max L$. Then $(\alpha_n, \sigma_{1,n}, \sigma_{2,n})$ are trapped in the compact set

$$\mathcal{C} = \{(\alpha, \sigma_1, \sigma_2) : L(\alpha_0, \sigma_{1,0}, \sigma_{2,0}) \le L(\alpha, \sigma_1, \sigma_2) \le L_\infty\}. \tag{13}$$

By (11), we have $\|L'\| \ge \delta > 0$ uniformly on $\mathcal{C}$. Fix $h_0 > 0$. Let M_0 be the max of $\|M''\|$ over points within h_0 of $\mathcal{C}$. Now, consider a maximization step. Suppose for instance we have just maximized on α, so we are at a point where $\partial L/\partial\alpha = 0$. Then

$$\Big(\frac{\partial L}{\partial \sigma_1}\Big)^2 + \Big(\frac{\partial L}{\partial \sigma_2}\Big)^2 \ge \delta^2. \tag{14}$$

We can make an h-step in the (σ_1, σ_2) direction, with $0 < h \le h_0$, increasing the log likelihood by at least

$$h\delta - \frac{1}{2} h^2 M_0 = \frac{\delta^2}{2M_0} > 0 \tag{15}$$

when $h = \delta/M_0$, provided $\delta/M_0 \le h_0$; or

$$h_0\delta - \frac{1}{2} h_0^2 M_0 > 0 \tag{16}$$

when $h = h_0$, provided $\delta/M_0 > h_0$. The maximization has to increase the log likelihood by even more. Thus, the log likelihood increases by some minimal positive amount

$$\eta = \min\Big\{\frac{\delta^2}{2M_0},\, h_0\delta - \frac{1}{2} h_0^2 M_0\Big\} > 0 \tag{17}$$

at each iteration. This is a contradiction, i.e, the log likelihood converges under iteration to its global maximum. Since the function is unimodal, the iterates must converge too. This completes the proof of Proposition 1.

We consider also convergence of the GLS algorithm when L is bimodal, under the side condition that the two modes differ. (Presumably, that excludes only a null set of samples.) If we exclude small open disks around the three critical points, and require $L \ge L_0$, the iteration can visit the resulting compact set only finitely often: otherwise the

log likelihood function would have an infinite maximum, as argued above. If the iteration comes arbitrarily close to the global maximum, convergence to that maximum is assured, since the log likelihood cannot decrease. Set that case aside: then the same reasoning applies to the second maximum, and finally to the saddle point. In short, the iterates must converge to one of the three critical points. Can the iterates converge to the saddle point? That depends on the local geometry, which we have not assessed.

Proposition 2. The log likelihood function $\mathfrak{L}$ is concave for no sample that satisfies condition (2).

Argument. We rewrite (3) with $\lambda_i = 1/\sigma_i^2$. Let $\mathfrak{L}$ be the log likelihood function (3) in this new parameterization. Then

$$-\frac{2}{n}\mathfrak{L}'(\alpha,\lambda_1,\lambda_2) = \begin{pmatrix} 2(\lambda_1+\lambda_2)\alpha - 2\lambda_1 y_1 - 2\lambda_2 y_2 \\ s_1^2 + (y_1-\alpha)^2 - 1/\lambda_1 \\ s_2^2 + (y_2-\alpha)^2 - 1/\lambda_2 \end{pmatrix}. \tag{18}$$

$$-\frac{2}{n}\mathfrak{L}''(\alpha,\lambda_1,\lambda_2) = \begin{pmatrix} 2(\lambda_1+\lambda_2) & 2(\alpha-y_1) & 2(\alpha-y_2) \\ 2(\alpha-y_1) & 1/\lambda_1^2 & 0 \\ 2(\alpha-y_2) & 0 & 1/\lambda_2^2 \end{pmatrix}. \tag{19}$$

The matrix in (19) is symmetric, so there are three real zeros in the characteristic polynomial

$$\big[2(\lambda_1+\lambda_2)-x\big]\big[\frac{1}{\lambda_1^2}-x\big]\big[\frac{1}{\lambda_2^2}-x\big]+4\big[(\alpha-y_1)^2+(\alpha-y_2)^2\big]x-4\frac{(\alpha-y_1)^2}{\lambda_2^2}-4\frac{(\alpha-y_2)^2}{\lambda_1^2}. \tag{20}$$

This cubic is by and large decreasing from ∞ at $x=-\infty$ to $-\infty$ at $x=\infty$, with a wiggle after it crosses the horizontal axis. There is a negative root provided the polynomial is negative at 0, i.e.,

$$\lambda_1+\lambda_2 < 2(\alpha-y_1)^2\lambda_1^2 + 2(\alpha-y_2)^2\lambda_2^2. \tag{21}$$

If the λ's are held fast, (21) will be satisfied for large α; also, if α is fixed, (21) will be satisfied for large λ's. (If $\alpha = y_1$ or y_2, a little care is needed in choosing the λ's.) Hence there is a negative eigenvalue for the matrix in (19). Thus, for any sample, there are regions in parameter space where the log likelihood function $\mathfrak{L}$ is not concave, i.e., the matrix of second derivatives has a positive eigenvalue. Signs may be confusing: (19) gives the second derivatives of $-\mathfrak{L}$. A negative eigenvalue for that matrix corresponds to a positive eigenvalue for $\mathfrak{L}$ itself, i.e., non-concavity. The argument for Proposition 2 is complete.

Remark. Of course, there remains the possibility that the log likelihood function will be everywhere concave in some other parameterization.

Example 2. Suppose the Y_{1j} are IID N(α, 1) and the Y_{2j} are IID N(α, σ^2) where α, σ are unknown. Apparently, the log likelihood function is unimodal but not convex in (α, $1/\sigma^2$). As before, the log likelihood function is $\mathcal{L}(\alpha, \lambda)$ where $\lambda = 1/\sigma^2$. Then

$$-2\mathcal{L} = (\alpha - y_1)^2 + \lambda[s^2 + (\alpha - y_2)^2] - \log \lambda$$

and

$$-2\mathcal{L}' = \begin{pmatrix} 2(\alpha - y_1) + 2\lambda(\alpha - y_2) \\ s^2 + (\alpha - y_2)^2 - 1/\lambda \end{pmatrix}.$$

Fix α and minimize in λ to get

$$M(\alpha) = (\alpha - y_1)^2 + \log\,[s^2 + (\alpha - y_2)^2]$$

with

$$\frac{1}{2}M'(\alpha) = \alpha - y_1 + \frac{\alpha - y_2}{s^2 + (\alpha - y_2)^2}$$

(The sign convention has reversed from the previous example.) By change of scale, put $y_1 = 1$ and $y_2 = 0$. Then $M'(\alpha) = 0$ iff

$$\alpha^3 - \alpha^2 + \alpha = s^2(1 - \alpha)$$

The left hand side is a strictly increasing function of α, with a zero at 0; the right hand side is strictly decreasing, with a zero at 1. Hence $M'(\alpha) = 0$ has exactly one root, and M has a unique minimum. Finally,

$$-\mathcal{L}'' = \begin{pmatrix} 1 + \lambda & \alpha - y_1 \\ \alpha - y_1 & \dfrac{1}{2\lambda^2} \end{pmatrix}$$

will not be positive definite if α is large, or λ is large.

Example 3. Suppose $X \sim \mathrm{N}(\alpha, 1)$ and $Y \sim \mathrm{N}(\beta, 1)$. The two random variables are independent. The parameters α, β are constrained to lie on a smooth curve $\mathcal{C}$. Unless $\mathcal{C}$ is a straight line, there will always be samples with a multimodal log likelihood function. These samples may be rather remote from the true value or the global max.

Example 4. Suppose $X \sim \mathrm{N}(\alpha, 1)$ and $Y \sim \mathrm{N}(\beta, 1)$. The two random variables are independent. The parameters α, β are constrained to lie on a smooth curve $\mathcal{C}$, to be constructed next. For any $\delta > 0$, for all samples (X, Y) within a δ-neighborhood of truth, the log likelihood function is multimodal. This hold for some large set of true values.

Construction. The curve is defined as follows: take the upper part of the vertical axis $\{(\alpha, \beta) : \alpha = 0,\ \beta > 1\}$, together with the line segments joining (i) $(0, 1)$ to $(-1, 0)$, then (ii) $(-1, 0)$ to $(1, 0)$, then (iii) $(1, 0)$ to $(0, -1)$, and finally the lower part of the vertical axis $\{(\alpha, \beta) : \alpha = 0,\ \beta < -1\}$. This curve can be smoothed to be C_∞ or even analytic, although in the latter case the line segments become curves. For any truth in the upper or lower vertical segments, for all (X, Y) inside any small disk around truth, the likelihood is multimodal. To smooth $\mathcal{C}$, we could view α as a function of β, then smooth this function using a tight gaussian kernel; the ambiguity at $\beta = 0$ in the definition of the unsmoothed function is immaterial.

Other examples. f is C_∞ on the plane, bounded above, tends to $-\infty$ at ∞.

(i) f is unimodal, but alternating maximization stagnates, and not at the maximum.
(ii) f has 3 critical values, with two modes and a saddlepoint. The two modes are of equal value. Alternating maximization oscillates between the two.

Construction. (i) Cut an orange in half. Put it down on the plane. Stand under the skin somewhere in the southwest quadrant, and push up the skin smoothly, without changing the other three quadrants. Keep going till the maximum is higher than the old central point, and make a smooth decline from the new max to the old max. If you start the iteration from anywhere in the other three quadrants, you will stabilize at the old max. Now deform the plane downward as you move away from the orange. With a little more effort, you can make the old max completely stable, i.e., the deformed orange has a flat spot there.

(ii) Cut a football in half, lengthwise. Put one half down centered at $(-1, 0)$ and the other half at $(1, 0)$. These are the two maxima, of equal height. Orient the football halves with the long axis at 45 degrees, towards the northeast. Thus, horizontal sections are congruent. Start the iteration from e.g. $(1.1, 0)$, assuming this is under the dome. Take the max on y. Then max on x: there is one on your football, and a matching spot on the other football: make a tiny pimple at the matching spot, so you go there. Proceed iteratively, with pimples getting smaller by leaps and bounds, and heights always below the old global maxima. Finally, deform the plane to go down as you move away from the domes.