Maximum likelihood estimation: actual or supposed

1. MLEs in exponential families

Let $f(x,\theta)$ for $x \in X$ and $\theta \in \Theta$ be a likelihood function, that is, for present purposes, either X is a Euclidean space $\mathbb{R}^d$ and for each $\theta \in \Theta$ $f(\cdot,\theta)$ is a probability density function on X, or X is a countable set, and $f(\cdot,\theta)$ is a probability mass function, or as Bickel and Doksum call it, a frequency function. Let P_θ be the probability distribution (measure) on X of which $f(\cdot,\theta)$ is the density or mass function. For the case of densities, let's assume that for each $\theta \in \Theta$, for each open set U of x on which $f(\cdot,\theta)$ equals almost everywhere a bounded continuous function, it equals that function everywhere on U.

For each $x \in X$, a *maximum likelihood estimate (MLE)* of θ is any $\widehat{\theta} = \widehat{\theta}(x)$ such that $f(\widehat{\theta},x) = \sup\{f(\phi,x) : \ \phi \in \Theta\} > 0$. In other words, $\widehat{\theta}(x)$ is a point at which $f(\cdot,x)$ attains its maximum and the maximum is strictly positive. In general, the supremum may not be attained, or it may be attained at more than one point. If it is attained at a unique point $\widehat{\theta}$, then $\widehat{\theta}$ is called *the* maximum likelihood estimate of θ. A measurable function $\widehat{\theta}(\cdot)$ defined on a measurable subset B of X is called a *maximum likelihood estimator* if for all $x \in B$, $\widehat{\theta}(x)$ is a maximum likelihood estimate of θ, and almost all x not in B, the supremum of $f(\cdot,x)$ is not attained at any point or is 0.

Define $W := \{x \in X : \ \sup_\theta f(\theta,x) = 0\}$. Very often, the set W will simply be empty. If it's non-empty and an $x \in W$ is observed, then there is no maximum likelihood estimate of θ. Moreover, for any prior π on Θ, a posterior distribution π_x can't be defined. That indicates that the assumed model $\{P_\theta\}_{\theta\in\Theta}$ is "misspecified," i.e. wrong, because according to the model, an observation in W shouldn't have occurred except with probability 0, no matter what θ is. Note that the set W is determined in advance of taking any observations. By contrast, for any continuous distribution on $\mathbb{R}$ say, each individual value has 0 probability, but we know only with hindsight (retrospectively) what the value is.

As is not surprising, a sufficient statistic is sufficient for finding MLEs:

Proposition 1. *For a family $\{P_\theta\}_{\theta\in\Theta}$ of the form described, suppose $T(x)$ is a sufficient statistic for the family. Except for x in a set B with $P_\theta(B) = 0$ for all θ, what values θ (none, one, or more) are MLEs of θ depend only on $T(x)$.*

Proof. This is a corollary of the factorization theorem for sufficient statistics. □

Remark. One would like to say that the set A of x for which an MLE exists and is unique is a measurable set and that on A, the MLE $\widehat{\theta}$ is (at least) a measurable function of x. Such statements are not generally true for Borel σ-algebras but may be true for larger σ-algebras such as that of analytic (also called Suslin) sets, e.g. Dudley (2002, Chapter 13). In practice, MLEs are usually found for likelihoods having derivatives, setting derivatives or gradients equal to 0 and checking side conditions. For example, for exponential families, MLEs, if they are in the interior of the natural parameter space, will be described in Theorem 3.

Example 2. (i) For each $\theta > 0$ let P_θ be the uniform distribution on $[0, \theta]$, with $f(\theta, x) := 1_{[0,\theta]}(x)/\theta$ for all x. Then if $X_1, \ldots, X_n$ are observed, i.i.d. (P_θ), the MLE of θ is $X_{(n)} := \max(X_1, \ldots, X_n)$. Note however that if the density had been defined as $1_{[0,\theta)}(x)$, the supremum for given $X_1, \ldots, X_n$ would not be attained at any θ. The MLE of θ is the smallest possible value of θ given the data, so it is not a very reasonable estimate in some ways. Given θ, the probability that $X_{(n)} > \theta - \delta$ approaches 0 as $\delta \downarrow 0$.

(ii) For $P_\theta = N(\theta, 1)^n$ on $\mathbb{R}^n$, with usual densities, the sample mean $\overline{X}$ is the MLE of θ. For $N(\mu, \sigma^2)^n$, $n \geq 2$, the MLE of (μ, σ^2) is $(\overline{X}, \sum_{j=1}^n (X_j - \overline{X})^2/n)$. Here recall that the usual, unbiased estimator of σ^2 has $n - 1$ in place of n, so that the MLE is biased, although the bias is small, of order $1/n^2$ as $n \to \infty$. The MLE of σ^2 fails to exist (or equals 0, if 0 were allowed as a value of σ^2) exactly on the event that all X_j are equal for $j \leq n$, which happens for $n = 1$, but only with probability 0 for $n \geq 2$. On this event, $f((\overline{X}, \sigma^2), x) \to +\infty$ as $\sigma \downarrow 0$.

(iii) In the previous two examples, for the usual choices of densities, the set W is empty, but here it will not be. Let $X = [0, +\infty)$ with usual Borel σ-algebra. Let $\Psi = (1, +\infty)$ and for $1 < \psi < \infty$ let Q_ψ be the gamma distribution with density $f(\psi, x) = x^{\psi-1}e^{-x}/\Gamma(\psi)$ for $0 \leq x < \infty$. Then $W = \{0\}$. If $x = 0$ is observed there is no MLE nor posterior distribution for ψ for any prior on Ψ.

In general, let Θ be an open subset of $\mathbb{R}^k$ and suppose $f(\theta, x)$ has first partial derivatives with respect to θ_j for $j = 1, \ldots, k$, forming the

gradient vector

$$\nabla_\theta f(\theta, x) \ := \ \{\partial f(\theta, x)/\partial \theta_j\}_{j=1}^k.$$

If the supremum is attained at a point in Θ, then the gradient there will be 0, in other words the *likelihood equations* hold,

$$\partial f(\theta, x)/\partial \theta_j \ = \ 0 \ \text{ for } \ j = 1, \ldots, k. \tag{1}$$

If the supremum is not attained on Θ, then it will be approached at a sequence of points $\theta^{(m)}$ approaching the boundary of Θ, or which may become unbounded if Θ is unbounded.

The equations (1) are sometimes called "maximum likelihood equations" in statistics books and papers, but that is unfortunate terminology because in general a solution of (1) could be (a) only a local, not a global maximum of the likelihood, (b) a local or global minimum of the likelihood, or (c) a saddle point, as in an example to be given in the next section.

For exponential families, it will be shown that an MLE in the interior of Θ, if it exists, is unique and can be found from the likelihood equations, as follows:

Theorem 3. *Let $\{P_\theta\}_{\theta \in \Theta}$ be an exponential family of order k, where Θ is the natural parameter space in a minimal representation. Let U be the interior of Θ and $j(\theta) \ := \ -\log C(\theta)$. Then for any n and observations $X_1, \ldots, X_n$ i.i.d. (P_θ), there is at most one MLE $\widehat{\theta}$ in U. The likelihood equations have the form*

$$\partial j/\partial \theta_i \ = \ \sum_{j=1}^n T_i(X_j)/n \ \ \textit{for} \ \ i = 1, \ldots, k, \tag{2}$$

and have at most one solution in U, which if it exists is the MLE. Conversely, any MLE in U must be a solution of the likelihood equations. If an MLE exists in U for v-almost all x, it is a sufficient statistic for θ.

Proof. Maximizing the likelihood is equivalent to maximizing its logarithm (the log likelihood), which is

$$\log f(\theta, x) = -nj(\theta) + \sum_{j=1}^n \sum_{i=1}^k \theta_i T_i(X_j),$$

and the gradient of the likelihood is 0 if and only if the gradient of the log likelihood is 0, which evidently gives the equations (2). Then $K = e^j$ is a smooth function of θ on U by Theorem 6 of the "Exponential Families" handout. hence so is j, and the other summand in the log

likelihood is linear in θ, so the log likelihood is a smooth (C^∞) function of θ. So at a maximum in U, the gradient must be 0, in other words (2) holds.

Recall that a real-valued function f on a convex set C is called *strictly convex* if for any $x \neq y$ in C and $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y).$$

A real-valued function f on a convex set in $\mathbb{R}^k$ is called *concave* if $-f$ is convex and *strictly concave* if $-f$ is strictly convex. It is easily seen that a strictly concave function on a convex open set has at most one local maximum, which then must be a strict absolute maximum. Adding a linear function preserves (strict) convexity or concavity. Now, j is strictly convex on U by Corollary 7 of "Exponential Families." So, if $\triangledown \log f(\theta, x) = 0$ at a point $\theta \in U$, then θ is a strict global maximum of $f(\cdot, x)$ as desired.

If for almost all x, (2) has a solution $\theta = \theta(x)$ in U, which must be unique, then by (2), the vector $\{\sum_{j=1}^n T_i(X_j)\}_{i=1}^k$, which is a sufficient k-dimensional statistic as noted in Theorem 1 of Exponential Families, is a function of $\theta(x)$ which thus must also be sufficient. □

Next is an example to show that if a maximum likelihood estimate exists almost surely but may be on the boundary of the parameter space, it may not be sufficient.

Proposition 4. *There exists an exponential family of order $k = 1$ such that for $n = 1$, a maximum likelihood estimate exists almost surely, is on the boundary of the natural parameter space with positive probability, and is not sufficient.*

Proof. Let the sample space X be $[1, \infty)$. Take the exponential family of order 1 having densities $C(\theta)e^{\theta x}/x^3$ (with respect to Lebesgue measure on $[1, \infty)$), where $C(\theta)$ as usual is the normalizing constant. Then the natural parameter space Θ is $(-\infty, 0]$, with interior $U = (-\infty, 0)$. We have $K(\theta) = \int_1^\infty e^{\theta x} x^{-3} dx$.

For $j(\theta) = \log K(\theta)$, we have by Corollary 7 of "Exponential Families" that $j''(\theta) > 0$ for $-\infty < \theta < 0$, so $j'(\theta)$ is increasing on U. We have

$$j'(\theta) = \frac{K'(\theta)}{K(\theta)} = \frac{\int_1^\infty e^{\theta x} x^{-2} dx}{\int_1^\infty e^{\theta x} x^{-3} dx}. \tag{3}$$

As $\theta \uparrow 0$, it follows by dominated convergence in the numerator and denominator that

$$j'(\theta) \uparrow \int_1^\infty x^{-2} dx \Big/ \int_1^\infty x^{-3} dx = 1/(1/2) = 2.$$

For $\theta \to -\infty$, multiply the numerator and denominator of the latter fraction in (3) by $|\theta|e^{-\theta}$. The law with density $|\theta|e^{\theta(x-1)}1_{[1,\infty)}(x)$ is that of $X+1$ where X is exponential with parameter $|\theta|$ and $EX = 1/|\theta|$. As $\theta \downarrow -\infty$, this distribution converges to a point mass at 1, and both functions x^{-2} and x^{-3} are bounded and continuous on $[1,\infty)$. Thus $j'(\theta) \downarrow 1$ as $\theta \downarrow -\infty$. So j' is increasing from $U = (-\infty, 0)$ onto $(1, 2)$. Hence for $1 < x < 2$, but not for $x \geq 2$, the likelihood equation (2) for $n = k = 1$ has a solution $\theta \in U$.

For $x \geq 2$, it will be shown that $f(\theta, x) = e^{\theta x}/(x^3 K(\theta))$ is maximized for $-\infty < \theta \leq 0$ at $\theta = 0$. As usual, maximizing the likelihood is equivalent to maximizing its logarithm. We have for $\theta < 0$ that $\partial \log f(\theta, x)/\partial\theta = x - j'(\theta) > 0$ since $x \geq 2 > j'(\theta)$ as shown above. Now $f(\theta, x)$ is continuous in θ at 0 from the left by dominated convergence, so for $x \geq 2$ it is indeed maximized at $\theta = 0$. Thus for $x \geq 2$ the MLE is $\hat{\theta} = 0$.

But, the identity function x is a minimal sufficient statistic by factorization. So the maximum likelihood estimator is not sufficient in this case although it is defined almost surely. The half-line $[2,\infty)$ has positive probability for each θ. Thus the proposition is proved. □

2. Likelihood equations and errors-in-variables regression; Solari's example

Here is a case, noted by Solari (1969), where the likelihood equations (1) have solutions, none of which are maximum likelihood estimates. It indicates that the method of estimation via "estimating equations," mentioned by Bickel and Doksum, should be used with caution.

Let (X_i, Y_i), $i = 1, \ldots, n$, be observed points in the plane. Suppose we want to do a form of "errors in variables" regression, in other words to fit the data by a straight line, assuming normal errors in both variables, so that $X_i = a_i + U_i$ and $Y_i = ba_i + V_i$ where $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are all jointly independent, with U_i having distribution $N(0, \sigma^2)$ and V_i distribution $N(0, \tau^2)$ for $i = 1, \ldots, n$. Here the unknown parameters are $a_1, \ldots, a_n$, b, σ^2 and τ^2. Let $c := \sigma^2$ and $h := \tau^2$. Then the joint density is

$$(ch)^{-n/2}(2\pi)^{-n} \exp\left(-\sum_{i=1}^{n} (x_i - a_i)^2/(2c) + (y_i - ba_i)^2/(2h) \right).$$

Let $\sum := \sum_{i=1}^{n}$. Taking logarithms, the likelihood equations are equivalent to the vanishing of the gradient (with respect to all $n+3$

parameters) of

$$-(n/2)\log(ch) - \sum[(X_i - a_i)^2/(2c) + (Y_i - ba_i)^2/(2h)].$$

Taking derivatives with respect to c and h gives

$$c \;=\; \sum(X_i - a_i)^2/n, \quad h \;=\; \sum(Y_i - ba_i)^2/n. \tag{4}$$

For each $i = 1, \ldots, n$, $\partial/\partial a_i$ gives

$$0 \;=\; c^{-1}(X_i - a_i) + h^{-1}b(Y_i - ba_i), \tag{5}$$

and $\partial/\partial b$ gives

$$\sum a_i(Y_i - ba_i) \;=\; 0. \tag{6}$$

Next, (5) implies

$$\sum(X_i - a_i)^2/c^2 \;=\; b^2\sum(Y_i - ba_i)^2/h^2. \tag{7}$$

From (4) it then follows that $1/c \;=\; b^2/h$ and $b \neq 0$. This and (5) imply that $b(X_i - a_i) \;=\; -(Y_i - ba_i)$, so

$$a_i \;=\; (X_i + b^{-1}Y_i)/2 \;\text{ for }\; i = 1, \ldots, n. \tag{8}$$

Then from (4) again,

$$c \;=\; \sum(X_i - b^{-1}Y_i)^2/(4n) \;\text{ and }\; h \;=\; \sum(Y_i - bX_i)^2/(4n). \tag{9}$$

Using (8) in (6) gives

$$\sum(Y_i - bX_i)(X_i + b^{-1}Y_i) \;=\; 0 \;=\; \sum Y_i^2 - b^2X_i^2.$$

If $\sum Y_i^2 > 0 = \sum X_i^2$ there is no solution for b. Also if $\sum Y_i^2 = 0 < \sum X_i^2$ we would get $b = 0$, a contradiction, so there is no solution in this case. If $\sum X_i^2 \;=\; \sum Y_i^2 \;=\; 0$ then (6) gives $a_i = 0$ for all i since $b \neq 0$, but then (4) gives $c = 0$, a contradiction, so there is no solution. We are left with the general case $\sum Y_i^2 > 0 < \sum X_i^2$. Then $b^2 \;=\; \sum Y_i^2/\sum X_i^2$, so

$$b \;=\; \pm\left(\sum Y_i^2/\sum X_i^2\right)^{1/2}. \tag{10}$$

Substituting each of these two possible values of b in (8) and (9) then determines values of all the other parameters, giving two points, distinct since $\sum Y_i^2 > 0$, where the likelihood equations hold, in other words *critical points* of the likelihood function, and there are no other critical points.

However, the joint density goes to $+\infty$ for $a_i = X_i$, fixed b and h, and $c \downarrow 0$. Thus the above two points cannot give an absolute maximum of the likelihood. On the other hand, as $c \downarrow 0$ for any fixed $a_i \neq X_i$,

the likelihood approaches 0. So the likelihood behaves pathologically in the neighborhood of points where $a_i = X_i$ for all i and $c = 0$, its logarithm having what is called an essential singularity. Other such singularities occur where $Y_i - ba_i \to 0$ and $h \downarrow 0$.

The family of densities in the example can be viewed as exponential, but in a special sense, where $x = ((X_1, Y_1), ..., (X_n, Y_n))$ is considered as just one observation. If we take the natural parameters for the family parameterized by $b, a_1, \ldots, a_n, \sigma^2, \tau^2$, we get not the full natural parameter space, but a curved submanifold in it. For example, let $n = 2$. If θ_i are the coefficients of X_i and θ_{i+2} those of Y_i for $i = 1, 2$, we have $\theta_1\theta_4 \equiv \theta_2\theta_3$. Also, for the natural parameters, some θ_j have σ^2 in the denominator, so that as $\sigma \downarrow 0$, these θ_j go to $\pm\infty$, where singular behavior is not so surprising.

Theorem 3 only guarantees uniqueness (not existence) of maximum likelihood estimates when they exist in the interior of the natural parameter space, and doesn't give us information about uniqueness, let alone existence, on curved submanifolds as here. In some examples given in problems, on the full natural parameter space, MLEs may not exist with positive probability. In the present case they exist with probability 0.

It seems that the model considered so far in this section is not a good one, in that the number of parameters $(n + 3)$ is too large and increases with n. It allows values of X_i to be fitted excessively well ("overfitted") by setting $a_i = X_i$. Alternatively, Y_i could be overfitted.

Having noted that maximum likelihood estimation doesn't work in the model given above, let's consider some other formulations.

Let Q_η, $\eta \in Y$ be a family of probability laws on a sample space X where Y is a parameter space. The function $\eta \mapsto Q_\eta$ is called *identifiable* if it's one-to-one, i.e. $Q_\eta \neq Q_\xi$ for $\eta \neq \xi$ in Y. If η is a vector, $\eta = (\eta_1, \ldots, \eta_k)$, a component parameter η_j will be called *identifiable* if laws Q_η with different values of η_j are always distinct. Thus, $\eta \mapsto Q_\eta$ is identifiable if and only if each component $\eta_1, \ldots, \eta_k$ is identifiable. Suppose $\theta \mapsto P_\theta$ for $\theta \in \Theta$ is identifiable and $Q_\eta \equiv P_{\theta(\eta)}$ for some function $\theta(\cdot)$ on Y. Then $\eta \mapsto Q_\eta$ is identifiable if and only if $\theta(\cdot)$ is one-to-one.

Example 5. Let $dQ_\eta(\psi) = ae^{\cos(\psi-\eta)}d\psi$ for $0 \leq \psi < 2\pi$ where a is the suitable constant, a subfamily of the von Mises–Fisher family. Then $\eta \mapsto Q_\eta$ is not identifiable for $\eta \in Y = \mathbb{R}$, but it is for $Y = [0, 2\pi)$ or $Y = [-\pi, \pi)$.

Now consider another form of errors-in-variables regression, where for $i = 1, \ldots, n$, $X_i = x_i + U_i$, $Y_i = a + bx_i + V_i$, $U_1, \ldots, U_n$ are i.i.d.

$N(0,\sigma^2)$ and independent of $V_1,\ldots,V_n$ i.i.d. $N(0,\tau^2)$, all independent of $x_1,\ldots,x_n$ i.i.d. $N(\mu,\zeta^2)$ where $a,b,\mu\in\mathbb{R}$ and $\sigma^2>0,\ \tau^2>0$ and $\zeta^2>0$. This differs from the formulation in the Solari example in that the x_i, now random variables, were parameters a_i in the example. In the present model, only the variables (X_i,Y_i) for $i=1,\ldots,n$ are observed and we want to estimate the parameters. Clearly the (X_i,Y_i) are i.i.d. and have a bivariate normal distribution. The means are $EX_i=\mu$ and $EY_i=\nu\ :=\ a+b\mu$. The parameters μ and ν are always identifiable. It is easily checked that $C_{11}\ :=\ \mathrm{Var}(X_i)\ =\ \zeta^2+\sigma^2$, $C_{22}\ :=\ \mathrm{Var}(Y_i)=b^2\zeta^2+\tau^2$, and $C_{12}=C_{21}\ :=\ \mathrm{Cov}(X_i,Y_i)\ =\ b\zeta^2$. A bivariate normal distribution is given by 5 real parameters, in this case $C_{11},C_{12},C_{22},\mu,\nu$. A continuous function (with polynomial components in this case) from an open set in $\mathbb{R}^6$ onto an open set in $\mathbb{R}^5$ can't be one-to-one by a theorem in topology on invariance of dimension (references are given to Appendix B of the 18.466 OCW 2003 notes), so the 6 parameters $a,b,\mu,\zeta^2,\sigma^2,\tau^2$ are not all identifiable.

If we change the problem so that $\lambda\ :=\ \tau^2/\sigma^2>0$ is assumed known, then all the 5 remaining parameters are identifiable (unless $\lambda C_{11}=C_{22}$), as follows. The equation for C_{22} now becomes $C_{22}=b^2\zeta^2+\sigma^2\lambda$. After some algebra, we get an equation quadratic in b,

$$b^2C_{12}+(\lambda C_{11}-C_{22})b-\lambda C_{12}\ =\ 0. \tag{11}$$

Since $\lambda>0$, the equation always has real solutions. If $C_{12}=0$ the equation becomes linear and either $b=0$ is the only solution or if $\lambda C_{11}-C_{22}=0$, b can have any value and in this special case is not identifiable. If $C_{12}\neq 0$ there are two distinct real roots for b, of opposite signs. Since b must be of the same sign as C_{12} to satisfy the original equations, there is a unique solution for b, and one can solve for all the parameters, so they are identifiable in this case.

Now, let's consider how the parameters can be estimated, in case λ is known.

Suppose given a normal distribution $N(\mu,C)$ on $\mathbb{R}^k$ where now $\mu=(\mu_1,\ldots,\mu_k)$, $C\ :=\ \{C_{ij}\}_{i,j=1}^k$ and $X_1,\ldots,X_n\in\mathbb{R}^k$ are n i.i.d. observations with $X_r\ :=\ \{X_{ri}\}_{i=1}^k$. It's easily seen that the MLE of μ is $\overline{X}\ :=\ \{\overline{X}_i\}_{i=1}^k$ where $\overline{X}_i\ :=\ \frac{1}{n}\sum_{r=1}^n X_{ri}$ for $i=1,\ldots,k$.

The classical unbiased estimator of the variance, called the *sample variance*, is defined for $n\geq 2$ as

$$s^2\ :=\ s_n^2\ :=\ \frac{1}{n-1}\sum_{j=1}^n(X_j-\overline{X})^2.$$

The MLE of the variance for a normal distribution is not s^2 but $s'^2 := (n-1)s^2/n$, which will be called the *empirical variance* here because it's the variance for the empirical probability measure P_n. Likewise, for a multivariate distribution, we can define sample covariances (with a factor $1/(n-1)$ in front) and empirical covariances (with a factor $1/n$). The latter again turn out to be MLEs for normal distributions:

Theorem 6. *The MLE of the covariance matrix C of a normal distribution on $\mathbb{R}^k$ is the empirical covariance matrix, for $i, j = 1, \ldots, k$,*

$$\hat{C}_{ij} := \frac{1}{n}\sum_{r=1}^{n}(X_{ri} - \overline{X}_i)(X_{rj} - \overline{X}_j).$$

Proof. For $k = 1$, this says that the MLE of the variance σ^2 for a normal distribution is $\frac{1}{n}\sum_{r=1}^{n}(X_r - \overline{X})^2$. (As noted above, this is $(n-1)/n$ times the usual, unbiased estimator s^2 of the variance.) This is easily checked, substituting in the MLE $\overline{X}$ of the mean, then finding that the likelihood equation for σ has a unique solution which is easily seen to give a maximum.

Now in k dimensions, consider any linear function f from $\mathbb{R}^k$ into $\mathbb{R}$ such as a coordinate, $f(X_r) = X_{ri}$. Let $\overline{f} := \frac{1}{n}\sum_{r=1}^{n} f(X_r)$. Then by the one-dimensional facts, $\overline{f}$ is the MLE of the mean of f and the MLE of $\mathrm{Var}(f)$ is the empirical variance $\frac{1}{n}\sum_{r=1}^{n}(f(X_r) - \overline{f})^2$.

For any function g on a parameter space, if a unique MLE $\widehat{\theta}$ of the parameter θ exists, then the MLE of $g(\theta)$ is, by definition, $g(\widehat{\theta})$. For example, if g is a one-to-one function, then $\eta = g(\theta)$ just gives an alternate parameterization of the family as $\mu_\eta = \mu_{g(\theta)} = P_\theta$. We see that the MLE of C_{jj} is $\hat{C}_{jj}$ for $j = 1, \ldots, k$. Moreover, the MLE of $\mathrm{Var}(X_{1i} + X_{1j})$ is also the corresponding empirical variance for any $i, j = 1, \ldots, k$. Subtracting the empirical variances of X_{1i} and X_{1j} and dividing by 2, we get the empirical covariance of X_{1i} and X_{1j}, namely $\hat{C}_{ij}$. Since the MLE of a sum or difference of functions of the parameters is the sum or difference of the MLEs, we get that the empirical covariance $\hat{C}_{ij}$ is indeed the MLE of C_{ij}. □

Returning to the bivariate case and continuing with errors-in-variables regression for a fixed λ, by a change of scale we can assume $\lambda = 1$, in other words $\sigma^2 = \tau^2$. Since $\overline{X}$ and $\overline{Y}$ are the MLEs of the means, we see that $\overline{Y} = \hat{\nu} = a + b\hat{\mu} = a + b\overline{X}$, the MLE regression line will pass through $(\overline{X}, \overline{Y})$, as it also does for the classical regression lines of y on x or x on y.

REFERENCE

Solari, Mary E. (1969). The "maximum likelihood solution" of the problem of estimating a linear functional relationship. *J. Roy. Statist. Soc.* **31**, 372-375.