More on nuisance parameters

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 3, Hilary Term 2009

January 30, 2009

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

C1: the distribution of C depends on λ but not on ψ;

C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;

C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \,|\, x) \propto f(s, c; \theta) = f(s \,|\, c; \psi) f(c; \lambda)$$

and we say that C is *ancillary for* ψ, S is *conditionally sufficient for* ψ given C, and C is *marginally sufficient for* λ.

We also say that C is a *cut* for λ and would then

- base inference about λ on the marginal distribution of C;

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

C1: the distribution of C depends on λ but not on ψ;

C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;

C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \,|\, x) \propto f(s, c; \theta) = f(s \,|\, c; \psi) f(c; \lambda)$$

and we say that C is *ancillary for* ψ, S is *conditionally sufficient for* ψ given C, and C is *marginally sufficient for* λ.

We also say that C is a *cut* for λ and would then

- base inference about λ on the marginal distribution of C;
- base inference about ψ on the conditional distribution of S given $C = c$.

Consider a sample $X = (X_1, \dots, X_n)$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Recall that $(U, V) = (\bar{X}, S^2 = \sum_i (X_i - \bar{X}_i)^2)$ is minimal sufficient and the likelihood function is

$$L(\mu, \sigma^2 \,|\, x) \propto f(u; \mu, \sigma^2) f(v; \sigma^2).$$

If we do straight maximum likelihood estimation, we have

$$\hat{\mu} = U = \bar{X}, \quad \hat{\sigma}^2 = V/n.$$

However, most statisticians agree that it is sensible to use $\tilde{\sigma}^2 = V/(n-1)$ as the estimator of σ^2. Is this reasonable and is there a general rationale for this?

Note that the *common unbiasedness argument does not work* as $\tilde{\sigma}$ is *not* unbiased for the standard deviation σ, or $\tilde{\sigma}^{-1}$ is *not* unbiased for the precision σ^{-2}.

This example shows that we have to be very careful when nuisance parameters are present and straight likelihood considerations can lead us astray:

We wish to establish the precision of a new instrument which measures with normal errors. We are therefore taking repeated measurements of individuals $(X_{i1}, X_{i2}), i = 1, \ldots, n$ which are all independent with

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2).$$

Now consider

$$U_i = (X_{i1} + X_{i2})/2, \quad V_i = (X_{i1} - X_{i2})/2.$$

These are again independent and normally distributed as

$$U_i \sim \mathcal{N}(\mu_i, \tau^2), \quad V_i \sim \mathcal{N}(0, \tau^2),$$

where $\tau^2 = \sigma^2/2$.

Clearly, we might as well consider (U_i, V_i) as the original data. Also, the pair (U, W) is minimal sufficient, where $U = (U_1, \ldots, U_n)$ and $W = \sum_i V_i^2$, hence the likelihood function becomes

$$\begin{aligned} L(\mu, \tau^2) &\propto (\tau^2)^{-n/2} e^{-\frac{1}{2\tau^2} \sum_i (u_i - \mu_i)^2} (\tau^2)^{-n/2} e^{-\frac{1}{2\tau^2} \sum_i v_i^2} \\ &= e^{-\frac{1}{2\tau^2} \sum_i (u_i - \mu_i)^2} (\tau^2)^{-n} e^{-\frac{w}{2\tau^2}}. \end{aligned}$$

Thus the maximum likelihood estimator is

$$\hat{\mu}_i = U_i, i = 1, \ldots, n; \quad \hat{\tau}^2 = W/2n.$$

But $W \sim \tau^2 \chi^2(n)$, so for large n, $\hat{\tau}^2 \approx n\tau^2/(2n) = \tau^2/2$!! So *the additional parameters μ_i are a serious nuisance if τ^2 is the parameter of interest*.

The previous example shows that straight likelihood considerations may not lead to meaningful results when only a part of the parameter is considered.

There are a number of suggestions for modifying the likelihood function to extract the evidence in the sample concerning a parameter of interest ψ when $\theta = (\psi, \lambda)$. Such modifications are generally known as *pseudo-likelihood* functions.

Examples include: *conditional* likelihood, *marginal* likelihood, *profile* likelihood, *integrated* likelihood, and others, for example local, partial, restricted, residual, penalized, etc. The many names bear witness that straight likelihood considerations may not always be satisfactory.

Suppose we can write the joint density of a sufficient statistic $T = (U, V)$ as

$$f(u; \lambda, \psi) f(v \mid u; \psi),$$

where ψ is the parameter of interest. Then, for fixed ψ, U is sufficient for λ. Inference for ψ can now be based on the *conditional likelihood function*

$$L(\psi; v \mid u) = f(v \mid u; \psi),$$

as the conditional distribution does not involve λ.

The critical issue is whether (useful) information about ψ is lost by ignoring the factor $f(u; \lambda, \psi)$.

In the normal example with many nuisance parameters, $U = (U_i, i = 1, \ldots, n)$ is sufficient for the nuisance parameter $\lambda = (\mu_i, i = 1, \ldots, n)$ for fixed $\psi = \tau^2$. Conditioning on U yields

$$L(\tau^2; w \mid u) = f(w \mid u; \tau^2) = f(w; \tau^2) = (\tau^2)^{-n/2} e^{-\frac{w}{2\tau^2}}.$$

This gives the conditional MLE $\hat{\tau}^2_{|u} = W/n$ which is more sensible.

It may be argued that $U_i \sim \mathcal{N}(\mu_i, \tau^2)$ cannot possibly have information about τ^2. Or at least that the information it may have is not useful.

Marginal likelihood uses conditioning the other way around. Suppose we can write the joint density of a sufficient statistic $T = (U, V)$ as

$$f(u \mid v; \lambda, \psi) f(v; \psi),$$

where ψ is the parameter of interest. Then the nuisance parameter λ can be eliminated by marginalization as it does not enter in the marginal distribution of V. Inference for ψ can now be based on the *marginal likelihood function*

$$L(\psi; v) = f(v; \psi).$$

The issue is also here whether (useful) information about ψ is lost by ignoring the factor $f(u \mid v; \lambda, \psi)$.

In the normal example with many nuisance parameters with $\lambda = (\mu_i, i = 1, \ldots, n)$ and $\psi = \tau^2$ we get

$$L(\tau^2; w) = f(w; \tau^2) = (\tau^2)^{-n/2} e^{-\frac{w}{2\tau^2}},$$

which in this case is identical to the conditional likelihood function considered earlier and hence $\hat{\tau}^2_w = W/n$.

Marginal likelihood is in this case also known as *residual likelihood* because it is based on the residuals

$$\begin{aligned} R_{i1} &= X_{i1} - \hat{\mu}_{i1} = X_{i1} - \frac{X_{i1} + X_{i2}}{2} = \frac{X_{i1} - X_{i2}}{2} = V_i \\ R_{i2} &= X_{i2} - \hat{\mu}_{i2} = X_{i2} - \frac{X_{i1} + X_{i2}}{2} = -V_i. \end{aligned}$$

The corresponding estimates are then known as *REML* estimates.

Marginal and conditional likelihood changes the problem either by ignoring some of the data (by marginalization) or by ignoring their variability (by conditioning).

Profile likelihood attempts to stick to the original data distribution and likelihood function, but eliminates the nuisance parameters by maximization.

The *profile likelihood function* $\hat{L}(\psi)$ for ψ is defined as

$$\hat{L}(\psi) = \sup_{\lambda} L(\psi, \lambda) = L\{\psi, \hat{\lambda}(\psi)\},$$

where ψ is the parameter of interest and $\hat{\lambda}(\psi)$ is the MLE of λ when ψ is considered fixed.

Although the profile likelihood generally can be very useful, it does not help in the the normal example with many nuisance parameters with $\lambda = (\mu_i, i = 1, \ldots, n)$ and $\psi = \tau^2$ we get

$$\hat{L}(\tau^2; w) = f(u; \hat{\mu}, \tau^2) f(w; \tau^2) = (\tau^2)^{-n} e^{-\frac{w}{2\tau^2}},$$

hence also peaks in the wrong place, at $\hat{\tau}^2 = W/(2n)$.

We shall later return to various attempts at modifying the profile likelihood.

Another way of removing nuisance parameters from the likelihood is to use integration. This method is essentially Bayesian and demands the specification of a prior distribution $\pi(\lambda \mid \psi)$ of the nuisance parameter for fixed ψ.

The *integrated likelihood function* is then defined as

$$\bar{L}(\psi) = \int L(\psi, \lambda)\pi(\lambda \mid \psi)\, d\lambda.$$

The integrated likelihood has the *same fundamental relation to the marginal prior and posterior distributions as the ordinary likelihood.*

For if $\pi(\psi)$ is the prior on ψ, the full posterior distribution is determined as

$$\pi^*(\psi, \lambda) \propto \pi(\psi)\pi(\lambda \mid \psi)L(\psi, \lambda)$$

and thus, by integration

$$\pi^*(\psi) \propto \int \pi^*(\psi, \lambda)\, d\lambda = \pi(\psi)\bar{L}(\psi).$$

In the normal example with many nuisance parameters, we may for example consider μ_i independent and normally distributed as $\mu_i \sim \mathcal{N}(\alpha, \omega^2)$, where (α, ω^2) represent prior knowledge about the population from which μ_i's are taken.

The integrated likelihood for τ^2 can then be calculated as

$$\bar{L}(\tau^2) = f(w; \tau^2) \int \prod_i f(u_i; \mu_i) \pi(\mu_i; \alpha, \omega^2)\, d\mu_i.$$

The integral can be recognized as the marginal distribution of U where now U_i are independent and identically distributed as $\mathcal{N}(\alpha, \tau^2 + \omega^2)$.

Thus

$$\begin{aligned}\bar{L}(\tau^2) &\propto f(w;\tau^2)(\tau^2+\omega^2)^{-n/2}e^{-\frac{1}{2(\tau^2+\omega^2)}\sum_i(u_i-\alpha)^2}\\ &\propto (\tau^2)^{-n/2}e^{-\frac{w}{2\tau^2}}(\tau^2+\omega^2)^{-n/2}e^{-\frac{q_\alpha(u)}{2(\tau^2+\omega^2)}}\end{aligned}$$

where

$$Q_\alpha(U)=\sum_i(U_i-\alpha)^2.$$

In this calculation, ω^2 and α are *known and fixed.* If these are 'correct', in the sense that μ_i are in fact behaving as if they were i.i.d. $\mathcal{N}(\alpha,\omega^2)$, then the integrated likelihood will peak around the correct value, else the peak will be shifted to an incorrect position. So the influence of the prior *prevails*.

Empirical Bayes or, equivalently(!), *MLE in the random effects model,* would also estimate α and ω^2 and get it right, as would *Hierarchical Bayes*, assigning a prior on (α,ω^2).