Mathematical Statistics
Stockholm University

Flat and multimodal likelihoods and model lack of fit in curved exponential families

Rolf Sundberg

Research Report 2009:1

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2009:1**,
http://www.math.su.se/matstat

Flat and multimodal likelihoods and model lack of fit in curved exponential families

Rolf Sundberg[*]

January 2009

Abstract

It is well-known that curved exponential families can have multimodal likelihoods. We investigate the relationship between flat or multimodal likelihoods and model lack of fit, the latter measured by the score (Rao) test statistic W_U of the curved model as embedded in the corresponding full model. We provide a formula for W_U, or a lower bound for it, when data yield a locally flat or convex likelihood (root of multiplicity > 1, terrace point, saddle point, local minimum). The formula is related to the statistical curvature of the model, and it depends on the amount of Fisher information. We use three models as examples, including the Behrens–Fisher model, to see how a flat likelihood etc. by itself can indicate a bad fit of the model. The results are much related (dual) to those of Efron [*Ann. Statist.* **6** (1978) 362–376].

Key words: nonunique MLE, likelihood equations with multiple roots, score test, Rao test, statistical curvature, Behrens–Fisher model, seemingly unrelated regressions, SUR

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se. Website: www.math.su.se/∼rolfs

1 Introduction

In full exponential families, log-likelihood functions necessarily have a unique maximum (or supremum), since they are concave functions in the canonical parameterization of the model. As a consequence, the likelihood equations have (at most) one root. Outside this class of models, likelihoods may be multimodal for some datasets or even all possible datasets, and the non-uniqueness can cause various types of problems.

It is well-known that curved exponential families can have multimodal likelihoods. Two important examples of such models are the Behrens–Fisher model (two samples from Gaussian distributions with the same mean value parameter but different variances), and the Seemingly Unrelated Regressions (SUR) model (bivariate or multivariate Gaussian linear regression with different sets of regressors for the different response variates — much used in econometrics). Characterized as a curved exponential family the Behrens–Fisher model is a (4, 3) model, meaning that the minimum sufficient statistic is of dimension 4 but the effective parameter dimension is only 3. A simple bivariate SUR model referred to below is (7, 5). In recent years, there has been increased interest in multimodality for these models types.

For the Behrens–Fisher model, Segiura and Gupta (1987) established that the likelihood equations can have several roots. Drton (2008) and Buot et al. (2007) developed probability asymptotics and generalizations to more than two populations, respectively. What happens in this situation is essentially that if the sample means are too distant apart, they will each yield a local maximum in their vicinity, and in between them is a saddle point more or less near the total average and representing what would approximately have been the unique maximum, had the sample means been less far apart.

The relatively simple (7, 5) bivariate SUR model, which has only one regressor for each response variable and zero intercepts, but an unknown residual covariance matrix, was investigated by Drton and Richardson (2004), who demonstrated that the log-likelihood can have several local maxima. This was an important observation, because it contradicted claims in the econometric literature. and it has consequences for MLE search algorithms. They also showed that under the SUR model this is a small sample phenomenon, because the probability for it to happen will go to zero with the sample size. On the other hand, the phenomenon can occur more frequently if the model is misspecified. Some results were extended by Drton (2005) to more complicated SUR models.

Simpler but more artificial examples that allow likelihood multimodality are found in Efron (1975, 1978), and in Barndorff–Nielsen and Cox (1994, Ch. 2). Example 2.38 in their book is a classical one, of a bivariate Gaussian distribution with the correlation coefficient as the only unknown parameter. This is a (2, 1) curved exponential family. Barndorff–Nielsen and Cox show by a diagram for what datasets multiple roots occur. An even simpler (2, 1) family is the "Normal parabola" (Example 2.35), two independent Gaussian variates with unknown means but known variances, and a specified parabolic relationship between the two means. These examples allow more explicit calculations, and are therefore important for checking and illustrating the theory.

In this paper the aim is not to obtain probabilistic statements about the occurrence of multiple roots under some specified model, but we will take a likelihood-based statistical view by assuming we are given data that yield a flat likelihood or multiple roots, and by investigating to what extent this relates to lack of fit of the curved model within a full model, in which it is embedded. We will see that the multimodality phenomenon is closely related to relatively large values of the score test statistic, to be denoted W_U. The results may be called dual in character to those of Efron's (1978) basic paper, by being likelihood-based rather than sampling-based in their interpretation. Also, Efron's paper is restricted to (2, 1) families, for conceptual simplicity.

2 Theory

We start with some basics about full and curved exponential families, see also Efron (1978) or Barndorff–Nielsen and Cox (1994, in particular sections 1.3 and 2.10). We set out from a basic model being a full exponential family, with likelihood function in canonical parameterization given as

$$\log L(\theta; t) = \theta^T t - \log(C(\theta)). \qquad (1)$$

Here $\theta \in \Theta$ is the canonical parameter, t is the canonical statistic (also minimum sufficient), and $C(\theta)$ is a norming constant which has the first and second order derivatives

$$\begin{aligned} D_\theta \log C(\theta) &= \mu_t(\theta) = E(t; \theta), & (2) \\ D_\theta^2 \log C(\theta) &= \mathrm{Var}(t; \theta). & (3) \end{aligned}$$

where μ_t is an alternative notation for the expectation vector of t. It follows that the Fisher score function $U(\theta; t)$ (or shorter $U(\theta)$) is the vector

$$U(\theta) = D_\theta \log L(\theta; t) = t - \mu_t(\theta), \tag{4}$$

and that the observed and the expected (Fisher) information matrices for θ are equal, being

$$-D_\theta^2 \log L(\theta) = -D_\theta U(\theta) = I(\theta) = \mathrm{Var}(t; \theta), \tag{5}$$

assumed nonsingular. The last relation shows that $\log L(\theta)$ is a concave function, hence it has a unique maximum, if any.

A curved exponential family is a subfamily specified by writing $\theta = \theta(\psi)$, where the subfamily parameter $\psi \in \Psi$ has a smaller dimension than the canonical parameter θ. We assume $\theta(\psi)$ is a smooth, intrinsically nonlinear function of ψ. The curved family has the same log-likelihood as the full family, only regarded as a function of $\theta(\psi)$, with ψ restricted to the space Ψ. Since $\theta(\psi)$ is nonlinear, the minimum sufficient statistic is still t, as for the full family.

The curved family score function $U(\psi)$ is obtained by using the chain rule,

$$U(\psi; t) = D_\psi \log L(\theta(\psi); t) = \left(\frac{\partial \theta}{\partial \psi}\right)^T (t - \mu_t(\theta(\psi))). \tag{6}$$

Here $\left(\frac{\partial \theta}{\partial \psi}\right)$ is the Jacobian matrix for the function $\theta(\psi)$. The expected information for ψ is

$$I(\psi) = \mathrm{Var}(U(\psi; t)) = \left(\frac{\partial \theta}{\partial \psi}\right)^T I(\theta(\psi)) \left(\frac{\partial \theta}{\partial \psi}\right). \tag{7}$$

The observed information differs from $I(\psi)$, being dependent on t, see below.

We first treat the relatively simple case $\dim \psi = 1$. This is enough to cover the last two examples mentioned in Section 1. They will reappear as Examples 3.1 and 3.2.

2.1 The case $\dim \psi = 1$.

The curved model $\theta = \theta(\psi)$ is here a one-dimensional curve in $\Theta \subset R^p$. The situation we have in mind is where the corresponding likelihood has a minimum or a very flat local or global maximum along the curve, around one or several roots $\hat{\psi}$ to the likelihood equation. We therefore assume that both the first and second derivatives of $\log L(\theta(\psi))$ are effectively zero at $\hat{\psi}$, or the latter even negative.

The second derivative of $\log L(\theta(\psi))$, or the observed information $J_t(\psi)$, is obtained by differentiating (6), with the scalar result

$$J_t(\psi) = -D^2_\psi \log L(\theta(\psi); t) = I(\psi) \; - \; \left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T (t - \mu_t(\theta(\psi))). \qquad (8)$$

For later use, note from (8) that

$$\mathrm{Var}(J_t(\psi); \psi) = \left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T I(\theta(\psi)) \left(\frac{\partial^2\theta}{\partial\psi^2}\right). \qquad (9)$$

When $J_t(\psi) \leq 0$, we have

$$\left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T (t - \mu_t(\theta(\psi))) \geq I(\psi) > 0. \qquad (10)$$

Furthermore, application of Cauchy–Schwarz inequality yields

$$\begin{aligned} & \left\{\left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T (t - \mu_t(\theta(\psi)))\right\}^2 \qquad (11) \\ \leq \; & \left\{\left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T I(\theta(\psi)) \left(\frac{\partial^2\theta}{\partial\psi^2}\right)\right\} \left\{(t - \mu_t(\theta(\psi)))^T I(\theta(\psi))^{-1} (t - \mu_t(\theta(\psi)))\right\}, \end{aligned}$$

where the first factor is $\mathrm{Var}(J_t(\psi))$, see (9). Combining (10) and (11) we get the inequality

$$(t - \mu_t(\theta(\psi)))^T I(\theta(\psi))^{-1} (t - \mu_t(\theta(\psi))) \geq \frac{I(\psi)^2}{\mathrm{Var}(J_t(\psi))}. \qquad (12)$$

When calculated in a MLE $\hat{\psi}$, the left hand side is the score (or Rao) test statistic W_U, for testing the fit of the curved model within the full model, see e.g. Barndorff–Nielsen and Cox (1994, ch. 3). The inequality yields a lower bound to the score test statistic. The bounds depends on the form of the curve and on the full model information matrix.

In (12), we have not yet utilized that ψ should be a root $\hat{\psi}$ of the likelihood equations. This means that we can sharpen (12). Because $\left(\frac{\partial\theta}{\partial\psi}\right)$ is orthogonal to $(t - \mu_t(\theta(\psi)))$ in a root $\hat{\psi}$, the left hand side of (10) does not change if the vector $\left(\frac{\partial^2\theta}{\partial\psi^2}\right)$ is replaced by its residual after projection on

$\left(\frac{\partial\theta}{\partial\psi}\right)$. As in (11) we use $I(\theta(\psi))$-norm, and then the projection is $b\left(\frac{\partial\theta}{\partial\psi}\right)$ with

$$b = \frac{\left(\frac{\partial^2\theta}{\partial\psi^2}\right)^T I(\theta(\psi))\left(\frac{\partial\theta}{\partial\psi}\right)}{\left(\frac{\partial\theta}{\partial\psi}\right)^T I(\theta(\psi))\left(\frac{\partial\theta}{\partial\psi}\right)}. \tag{13}$$

Thus, we obtain the sharper inequality

$$W_U(\hat{\psi}) \geq \frac{I(\hat{\psi})^2}{\left\{\left(\frac{\partial^2\theta}{\partial\psi^2}\right) - b\left(\frac{\partial\theta}{\partial\psi}\right)\right\}^T I(\theta(\hat{\psi}))\left\{\left(\frac{\partial^2\theta}{\partial\psi^2}\right) - b\left(\frac{\partial\theta}{\partial\psi}\right)\right\}}. \tag{14}$$

The right hand side is the squared radius of statistical curvature, as introduced by Efron (1975). Thus, we can equivalently write

$$W_U(\hat{\psi}) \geq 1/\gamma_{\hat{\psi}}^2, \tag{15}$$

where γ_ψ is Efron's statistical curvature (the inverse of the radius). An alternative expression for the right hand side, based on the fact that both $J(\psi)$ and $U(\psi)$ are linear forms in t, is

$$W_U(\hat{\psi}) \geq \frac{I(\hat{\psi})^2}{\mathrm{ResVar}(J|U;\hat{\psi})}, \tag{16}$$

where the denominator should be interpreted as the residual variance in a linear regression of J on U.

Remark 1. Note that if the information $I(\theta(\psi))$ is increased by a scalar factor, for example by increasing sample size, the lower bound in (14) will increase by the same factor. This means that as information accumulates, the occurrence of a multiple root or several roots along a flat likelihood will be a successively stronger indication that the model does not fit.

Remark 2. All t-vectors having a root $\hat{\psi}$ in common are situated in a $(\dim\theta - 1)$-dimensional hyperplane orthogonal to $\left(\frac{\partial\theta}{\partial\psi}\right)$. So is also the residual vector $\left(\frac{\partial^2\theta}{\partial\psi^2}\right) - b\left(\frac{\partial\theta}{\partial\psi}\right)$, with b from (13). When $\dim\theta = 2$, there is only one such direction, and inequalities (14–16) become equalities. In this case, the result is equivalent with Theorem 2 of Efron (1978). The differences are in the proof and in the interpretations. For $\dim\theta > 2$, equality is only attained for t in some particular direction from the point $\mu_t(\theta(\hat{\psi}))$, so in this case, there is only exceptional equality in (14–16).

2.2 The case $\dim \psi > 1$.

This case is somewhat more difficult to describe than the case $\dim \psi = 1$, but of interest to include not only for completeness, but because most curved families of applied importance have $\dim \psi > 1$. The Behrens–Fisher model will be investigated in Example 3.3 further below.

The observed information matrix $J_t(\psi)$ is positive definite precisely when the scalar $a^T J_t(\psi) a > 0$ for all vectors $a \neq 0$ (all directions). We can write

$$a^T J_t(\psi) a = a^T I(\psi) a - \left\{ a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a \right\}^T \{ t - \mu(\theta(\psi)) \},$$

where $a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a$ should be interpreted as a vector with elements $a^T \left(\frac{\partial^2 \theta_i}{\partial \psi^2} \right) a$. Continuing as in the case $\dim \psi = 1$, we see that the likelihood $L(\psi)$ has a minimum or a root of multiplicity > 1 in the a direction through $\hat{\psi}$ when

$$\left\{ a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a \right\}^T \left\{ t - \mu(\theta(\hat{\psi})) \right\} \geq a^T I(\hat{\psi}) a.$$

Applying Cauchy–Schwarz inequality and noting that the bound should hold whatever be a, we obtain the inequality

$$W_U(\hat{\psi}) \geq \min_{a \neq 0} \frac{\{a^T I(\hat{\psi}) a\}^2}{\left\{ a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a \right\}^T I(\theta(\hat{\psi})) \left\{ a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a \right\}}. \tag{17}$$

Thus, this lower bound holds for W_U whenever $\log L(\psi)$ is locally flat or locally convex in any direction a in Ψ through $\hat{\psi}$, e.g. if $\hat{\psi}$ is a local minimum, saddle point, or root of multiplicity > 1. This is exemplified and illustrated in Section 3.3.

As in Section 2.1, Remark 1, if the information matrix $I(\theta)$ is increased by a scalar factor, e.g. by an increased sample size, the lower bound is increased by the same factor.

In Section 2.1 the lower bound (12) could be sharpened by replacing $\frac{\partial^2 \theta}{\partial \psi^2}$ by a residual after projection on $\frac{\partial \theta}{\partial \psi}$. The corresponding replacement is not possible in the present case, even though the vector character of $a^T \frac{\partial^2 \theta}{\partial \psi^2} a$ might make us believe in full analogy. The reason is found by reconsidering Remark 2 of Sec. 2.1. If we think in terms of a one-dimensional submodel of Ψ, the residual vector is situated in a $(\dim \theta - 1)$-dimensional space, but not necessarily in the original $(\dim \theta - \dim \psi)$-dimensional space. Thus, the analogy fails.

3 Examples

3.1 Normal parabola, (2, 1).

Our first, simple model has been used as example also by Efron (1975) and Barndorff–Nielsen and Cox (1994, Example 2.35). Let y_1 and y_2 be two independent, unit variance Gaussian variates, with means unknown but following a simple parabolic relationship, $N_2(\psi, \psi^2, 1, 1, 0)$. This is a curved subfamily of the full model $N_2(\theta_1, \theta_2, 1, 1, 0)$, with canonical statistic $t = (y_1, y_2)^T$ and canonical parameter $\theta = (\theta_1, \theta_2)^T$. The full model score vector is $t - \theta$, the MLE is $\hat{\theta} = t$, and $I(\theta)$ is the identity matrix, making calculations simple. The curved model likelihood equation is

$$(y_1 - \psi) + 2\psi\,(y_2 - \psi^2) = 0, \tag{18}$$

which is a cubic equation in ψ that can have three real roots. However, only one root is of the same sign as y_1 (when $y_1 \neq 0$).

When $y_1 = 0$, there are three different roots as soon as $y_2 > 0.5$. The boundary case $y_2 = 0.5$ has a triple root (flat maximum) $\hat{\psi} = 0$, and the likelihood contours are shown in Figure 1a. The score test value is $W_U = 0.25$, and $1/\gamma^2_{\hat{\psi}}$ has the same value (see (15) and Remark 2). The value $W_U = 0.25$ is quite small, so there is nothing remarkable with a triple root or three roots along a relatively flat likelihood in this case.

For $y_1^2 \geq 2$, there are three real roots even when data fit the model perfectly. In that case, the global maximum of course corresponds to $W_U = 0$, and the lower bound (15) represents a more or less flat region around the local minimum. This is illustrated in Figure 1b, which represents the boundary case $y_1^2 = 2$, when the other two roots form a terrace point, so there is equality in (15). In the terrace point $\hat{\psi} = -1/\sqrt{2}$, $W_U = 6.75$, which is a high value. The conclusion is of course not that the whole model is wrong, only that the root $\hat{\psi} = -1/\sqrt{2}$ is an artefact of the curved model.

As stressed in Remark 1 above, the situation would be the same if we had a sample of n such pairs of data, only that we use $t = (\sum y_{1i}, \sum y_{2i})^T$. Provided the value of t/n remained the same as for $n = 1$, the score test statistic W_U and its lower bound would both be increased by the factor n. Thus, with a sample of size $n = 20$, say, a triple root $\hat{\psi} = 0$ would indicate a bad model fit.

3.2 Correlation example, (2, 1).

The setting here is a sample of size n from a marginally standardized bivariate Gaussian distribution, thus with only the correlation coefficient unknown, $N_2(0, 0, 1, 1, \rho)$. This is a curved subfamily of the full family $N(0, 0, \sigma^2, \sigma^2, \rho)$. If the latter is characterized by the canonical statistic

$$t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \begin{pmatrix} \sum(x_i^2 + y_i^2)/2 \\ \sum x_i y_i \end{pmatrix}$$

and the canonical parameter vector

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} -1/(1-\rho^2)\sigma^2 \\ \rho/(1-\rho^2)\sigma^2 \end{pmatrix},$$

the restricted model $\sigma = 1$ corresponds to the curve $\theta_1 = -0.5(1 + \sqrt{1 + 4\theta_2^2}$ in the cone Θ $(|\theta_2| \leq \theta_1)$. The expected value of t is given by $E(t_1/n) = \sigma^2$ and $E(t_2/n) = \rho\sigma^2$. The fact that the former is free of ρ does not mean that t_1 is ancillary for ρ, because the variance of t_1 depends on ρ. More precisely,

$$I(\theta) = n\sigma^4 \begin{pmatrix} 1+\rho^2 & 2\rho \\ 2\rho & 1+\rho^2 \end{pmatrix}$$

When $t_2 = 0$ is observed, both models have a root $\hat{\rho} = 0$ of the likelihood equation(s), whatever the value of $t_1 > 0$. In the full model it is of course the unique root and it represents a global maximum, but in the curved model this is not always true. If $t_1/n \geq 0.5$, the root is unique, but if $t_1/n < 0.5$ the root $\hat{\rho} = 0$ represents a local minimum, in between two identical maximum points. This is illustrated by Figures 2 and 3, which represent $t_1/n = 0.5$ and $t_1/n = 0.25 < 0.5$, respectively, and show the log-likelihood contours. In the limiting case $t_1/n = 0.5$, the root $\hat{\rho} = 0$ is a triple root, and the curve and the likelihood contour are seen to follow each other well around this point (Figure 2).

When $t_1/n = 0.25$, $t_2 = 0$ (Figure 3), the maximum is attained at $\rho \approx \pm 0.7$. A heuristic interpretation goes as follows. Consider $(t_1 \pm t_2)/n$, which are uncorrelated with expected values $1 \pm \rho$. When they have the same, relatively low, observed value 0.25, a conflict is created. The likelihood has a slight preference to go for a high positive or high negative ρ-value, fitting one of the expected values (at the expense of the other), rather than selecting the average position that would have been the maximum, had t_1 been 0.5 instead of 0.25.

When $t_2 \neq 0$, no triple root is possible. Figure 4 represents $t_1/n = 0.25$, $t_2/n = 0.118$. so the empirical correlation is positive, $r = 0.47$. The curved model has its likelihood maximum for $\rho = 0.86$ and a terrace point for $\rho = -0.37$. In the terrace point, W_U equals the lower bound, and the value is $W_U = 0.49\,n$.

3.3 Behrens–Fisher model, (4, 3).

The final example hsd $\dim \psi > 1$. For simplicity we let the two samples be of the same size, n, so in the full model we regard data as coming from $x_i \sim N(\mu_x, \sigma_x^2)$, $y_i \sim N(\mu_y, \sigma_y^2)$, $i = 1, ..., n$, with full mutual independence. The Behrens–Fisher model is obtained by specifying $\mu_x = \mu_y (= \mu)$. The problem is location and scale invariant, but we will additionally assume that we have observed the same sample variances. Then we can take $\bar{x} + \bar{y} = 0$ and $s_x^2 = s_y^2 = 1$ (sample variances with denominator n, not $n-1$), and only one statistic remains to be specified, the mean value difference $\bar{x} - \bar{y}$. From the symmetry of the setting we are led to think that $\hat{\mu} = 0$, $\hat{\sigma}_x = \hat{\sigma}_y = 1$ is the MLE, and, yes, there is such a root to the likelihood equation. However, it corresponds to a likelihood maximum only if $\bar{x}$ and $\bar{y}$ are not too wide apart, more precisely if $|\bar{x} - \bar{y}| \leq \sqrt{2}$. Otherwise it will be a local minimum symmetrically located between two maxima. The boundary case $|\bar{x} - \bar{y}| = \sqrt{2}$ is illustrated in Figure 5, which shows the two-dimensional profile log-likelihood for (μ_x, μ_y) and its restriction to the line $\mu_x = \mu_y$. The local flatness of the log-likelihood, for the common μ along that line, is striking.

The full model has $t_1 = n\bar{x}$, $t_2 = n\bar{y}$, $t_3 = n\overline{x^2}$, $t_4 = n\overline{y^2}$, and $\theta_1 = \mu_x/\sigma_x^2$, $\theta_2 = \mu_y/\sigma_y^2$, $\theta_3 = -0.5/\sigma_x^2$ $\theta_4 = -0.5/\sigma_y^2$. The restricted model has $\psi_1 = \mu$ (the common value of μ_x and μ_y), $\psi_2 = \theta_3$, and $\psi_3 = \theta_4$. Trivial calculations under $\hat{\psi}_1 = 0$ yield $\hat{\sigma}_x^2 = \overline{x^2}$ and $\hat{\sigma}_y^2 = \overline{y^2}$, with information matrices

$$I(\theta(\hat{\psi})) = n \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \qquad I(\hat{\psi}) = 2n \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence, with $|a| = 1$, the numerator of (17) is $(2n\,a^T a)^2 = 4\,n^2$. The denominator is found from

$$\left\{ a^T \left(\frac{\partial^2 \theta}{\partial \psi^2} \right) a \right\}^T = -4 a_1\,(a_2\; a_3\; 0\; 0)$$

to be $16\,n\,a_1^2\,(a_2^2+a_3^2) = 16\,n\,a_1^2\,(1-a_1^2)$ with maximum $4n$. Thus, the lower bound (17) is simply n.

Furthermore,

$$W_U(\hat{\psi}) = (n\bar{x}\ n\bar{y}\ 0\ 0)\,I(\theta(\hat{\psi}))^{-1}\,(n\bar{x}\ n\bar{y}\ 0\ 0)^T = n\,(\bar{x}^2+\bar{y}^2) = 2n\bar{x}^2.$$

Now, as remarked above, we have a root of multiplicity 3 precisely when $\bar{x} = 1/\sqrt{2}$, for which $W_U = n$, so W_U actually equals its lower bound (17).

4 Discussion

We have here provided a lower bound for the score test statistic $W_U(\hat{\psi})$ for model lack of fit test. The bound was derived for locally flat likelihoods, or for $\hat{\psi}$ being a minimum or saddle point. In three examples, the bound was calculated explicitely and related with the form of the likelihood surface.

The specific value of the bound is not of much practical interest, however, even though it is best possible, because when faced with data we can compute W_U for any $\hat{\psi}$. The most important features of the bound are theoretical: that it exists and that it increases proportionally to Fisher information (e.g. sample size). This is more explicitly expressed in the case $\dim\psi = 1$, when the lower bound is the squared radius of statistical curvature. This means that when sample size is large, a flat likelihood necessarily implies that the model does not fit data, whereas this need not be true for small data sets.

It is important to note the direction of the conclusions. That a model does not fit the data is not by itself a reason to expect multimodal or flat likelihoods. Typically deviations from the model lead to multimodality when t is on one side of the curved model, but not on the other, for example in the correlation example, Example 3.2. Formula (8) shows that the observed information will be greater than the expected on one side, but smaller on the opposite side.

We may think of a flat likelihood, as illustrated in Figures 2a, 2 and 5, as a boundary case between a unique maximum and three roots of which the middle is a local minimum or saddle point. In all three examples we saw that by gradually modifying data, we can make the same $\hat{\psi}$ change from a unique maximum to a minimum or saddle point (Figures 2 and 3). In Figure 3 the "natural" correlation estimate $\hat{\rho} = 0$ has been taken over as MLE by parameter values on each side of it, which are slightly less implausible under the model stated.

One conclusion is that a local minimum or saddle point might often be the “right” estimate under a wider, more reasonable model for data. So if data yield three roots of similar log-likelihood values, question the model.

Cox (2006, sec. 7.1) says about multimodal likelihoods that it may happen that there are two or more local maxima of similar size, but that “the more common situation is that the global maximum is dominant”. The latter situation corresponds to the first of the following three scenarios. The others are perhaps less common, but they are common enough to be of concern, in particular in connection with misspecified models.

Here are three typical scenarios to be distinguished:

- The global maximum is pronounced, but the likelihood is flat in another part of the parameter space, and for that reason may exhibit multiple roots. Parameter values in the flat region are likely to yield relatively large W_U-values, cf. Example 3.1.

- The likelihood is flat in some direction around its global maximum, and may exhibit multiple roots of the likelihood equation in or near that point. We have here provided a formula (or lower bound) for W_U, showing that this need not be a remarkable feature for small samples, but for larger samples it will indicate that the model does not fit the data. This was illustrated in all three examples.

- The likelihood has two widely separated maxima of similar magnitude, and a saddle point or minimum in between. Examples 3.2 (Fig. 3) and 3.2 (Fig. 5) illustrated that a likelihood can be essentially flat over a quite wide region, when the model is bad. When the likelihood is not that flat, the present results are not immediately applicable in the maximum points. However, it is still recommended to check the fit of the model.

Acknowledgement

Thanks are due to Nanny Wermuth, whose questions initiated this study, and whose interest in the SUR model might lead to a follow-up paper.

References

Barndorff–Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

Buot, M.-L.G., Hosten, S. and Richards, D.St.P. (2007). Counting and locating the solutions of polynomial systems of maximum likelihood equtions II: The Behrens–Fisher problem. *Stat. Sinica* **17**, 1343–1354.

Cox, D.R. (2006). *Principles of Statistical Inference.* Cambridge Univ. Press, Cambridge.

Drton, M. (2005). Computing all roots of the likelihood equations of seemingly unrelated regressions. *J. Symbolic Computation* **41**, 245–254.

Drton, M. (2008). Multiple solutions to the likelihood equations in the Behrens–Fisher problem. *Statistics and Probability Letters* **78**, 3288–3293.

Drton, M. and Richardson, T.S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91**, 383–392.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.* **3**, 1189–1242.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6**, 362–376.

Sugiura, N. and Gupta, A.K. (1987). Maximum likelihood estimates for Behrens–Fisher problem. *J. Japan Statist. Soc.* **17**, 55–60.

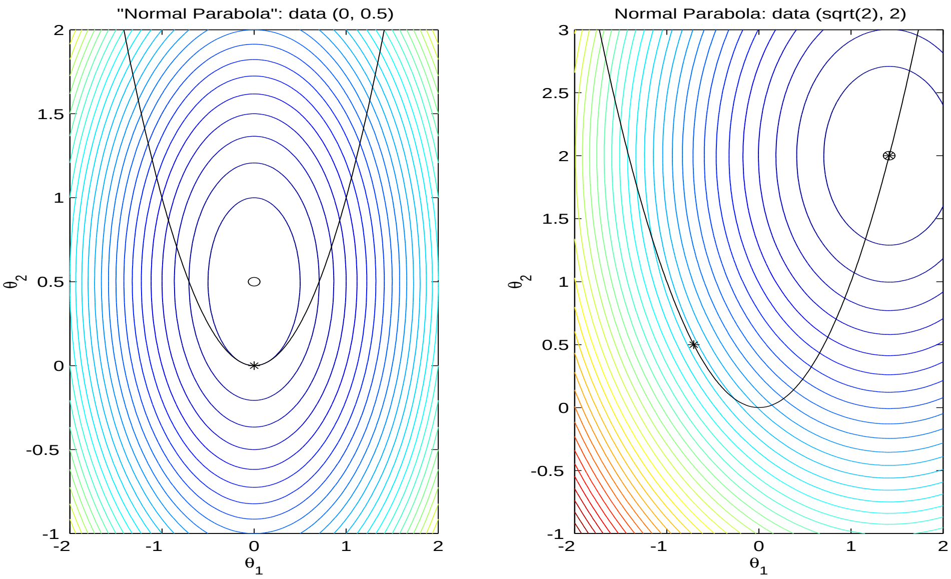

Figure 1: Normal parabola example.
a) Log-likelihood contours when $y_1 = 0, y_2 = 0.5$. Curved parameter relationship marked, and the triple root.
b) Log-likelihood contours when $y_2 = y_1^2 = 2$. Curved parameter relationship marked, and the two extremal points.

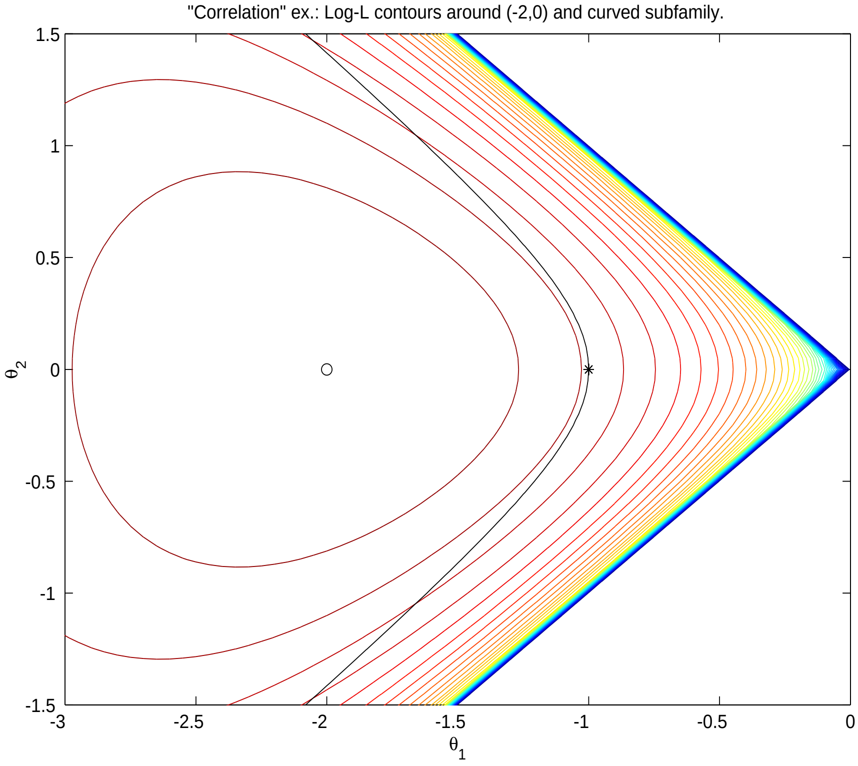

Figure 2: Correlation example with flat likelihood.
Log-likelihood contours around unconstrained maximum when $t_1/n = 0.5$, $t_2/n = 0$. Curved relationship and triple root $\hat{\rho} = 0$ marked.

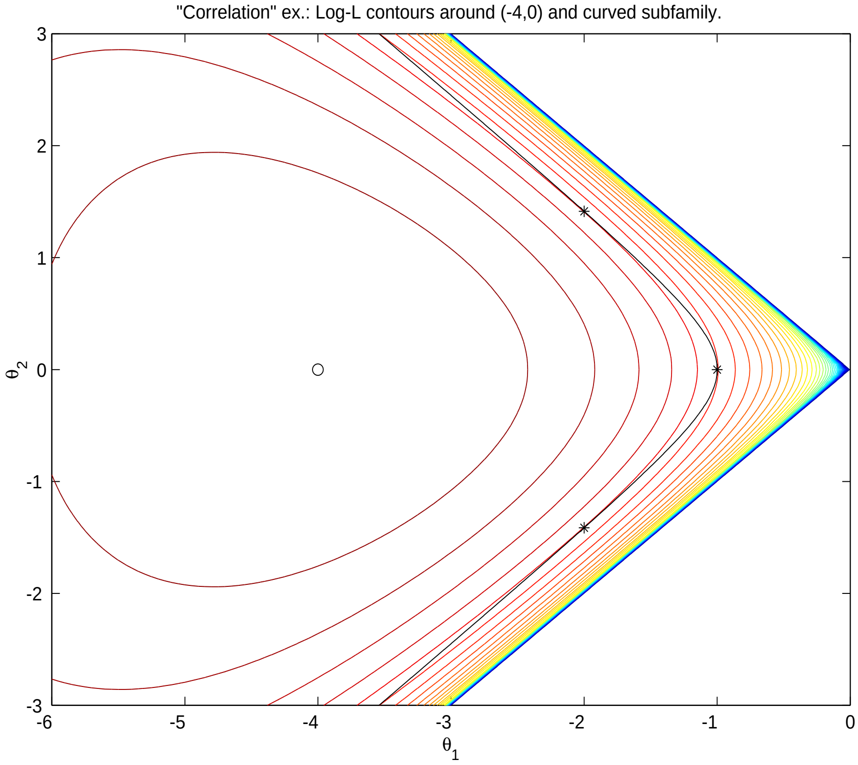

Figure 3: Correlation example with almost flat likelihood.
Log-likelihood contours around unconstrained maximum when $t_1/n = 0.25$, $t_2/n = 0$. Curved relationship and the three different roots marked ($\hat{\rho} = 0$ and $\hat{\rho} \approx \pm 0.7$).

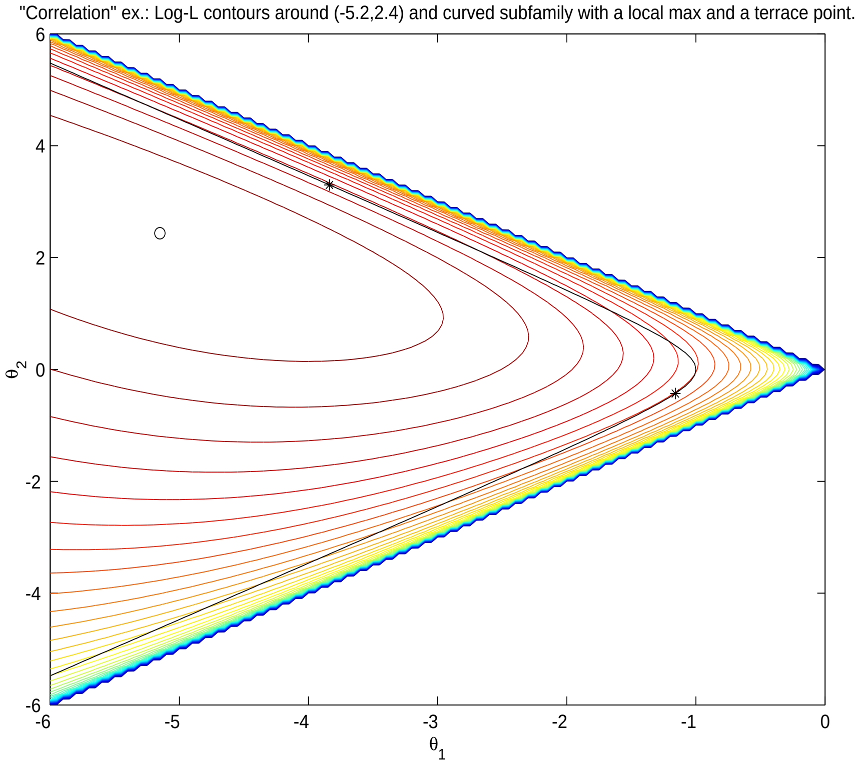

Figure 4: Correlation example with unsymmetrical roots.
Log-likelihood contours around unconstrained maximum when $t_1/n = 0.25$, $t_2/n = 0.118$. Curved relationship, root $\hat{\rho} = 0.86$ and double root $\hat{\rho} = -0.37$ marked.

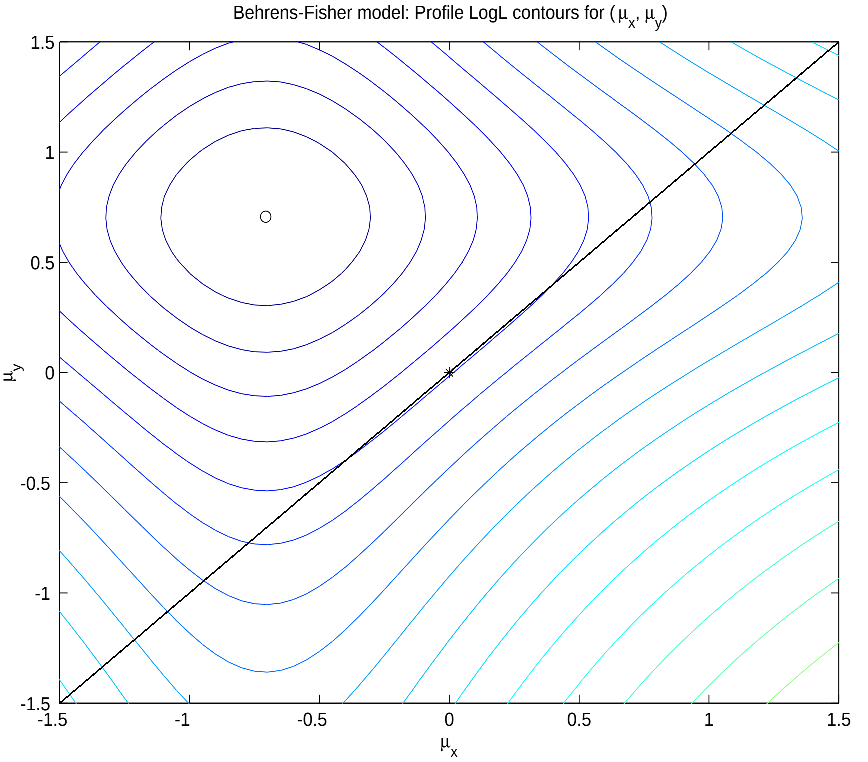

Figure 5: Behrens–Fisher model with flat likelihood maximum (triple root). The figure shows two-dimensional profile log-likelihood contours for (μ_x, μ_y) and for common μ (along line $\mu_x = \mu_y$), when sample sizes are equal, sample variances happen to be equal, and sample means differ precisely much enough to yield multiple roots. The circle indicates the observed mean values, and the fitted MLE is marked by an asterisk