Likelihood as Concept of Uncertainty

Marius Pfeuffer

21.04.2014

Supervisor: Dr. Marco Cattaneo
Institute of Statistics
Ludwig-Maximilians-Universität München

Abstract

The likelihood concept is a practically most relevant framework for statistical inference. It can be used to fit models to given data, derive point and interval estimates for parameters or conduct testing, to name possible fields of application. Within this light, the focus of this work is to systematically introduce the likelihood concept, point out the related methods for statistical inference and show shortcomings and anomalous behaviour. Inference is also discussed in the context of multivariate likelihood and possibilites for dealing with nuisance parameters are shown. Lastly, the likelihood concept is compared to the Bayesian and frequentist approach and it is shown how the likelihood concept is discussed in the scientific literature.

Contents

1 Motivation

Without any doubt, the likelihood concept is a comprehensive inference approach of major practical relevance. It can be used to fit models to given data, derive point and interval estimates for parameters or conduct testing, to name some of the fields of application. Within this light, the aim of this work is threefold:
Firstly, section 2 aims to systematically introduce the likelihood concept: its underlying basic principles and its methods for statistical inference. Furtheremore, this section shows examples for critical or anomalous behaviour of the likelihood function and the strong likelihood principle. In section 3, inference is discussed within the context of multivariate likelihood and possibilites for dealing with nuisance parameters are shown and critically discussed.
Secondly, section 4 deals with the history of the likelihood concept as well as how it can be compared to the Bayesian or frequentist approach. Additionally, it is shown how the likelihood concept is discussed in scientific literature.
Thirdly, throughout this work, contributions and hints from the discussion of the related seminar "Probability and Other Concepts of Uncertainty" which took place in March 2014 at the University of Munich are also used. To highlight these discussion related text passages, they are marked with [*].
This work takes especially into account sources by Pawitan [16] and Edwards [6].

2 The Likelihood Concept

2.1 Likelihood and Uncertainty

Both, probability and likelihood are concepts for dealing with uncertainty. Probability quantifies uncertainty of outcomes of random variables. Hence, when probability is described by parametric models, it quantifies the uncertainty of outcomes of random variables given the parameters of the assumed probability model.
This provides the link to the likelihood approach. When the parameters of such models are not known, but realizations of the corresponding random variables are available, the likelihood function quantifies the uncertainty of the parameters taking particular values, given the observed realizations.
In a Bayesian context, probability in terms of prior and posterior distributions is different.

The "prior probability" quantifies the previous knowledge about the uncertainty of the parameter, the "posterior probability" quantifies the uncertainty of the parameter given the prior knowledge and the observed realizations.

The foundation of the likelihood concept is the likelihood function. Below, the likelihood function is defined more formally. Hence, the link between likelihood and parametric probability models can be seen more clearly.

Likelihood Function

The likelihood function is a function of an unobserved parameter (here: λ) given a sample (here: x).

$$L(\lambda|x) = \mathbb{P}(x|\lambda)$$

The unobserved parameter λ refers to a probability model $\mathbb{P}(x|\lambda)$ which is assumed for the data x. This underlying probability model can be either discrete or continuous. In the discrete case, the likelihood function is defined by a discrete probability function

$$L(\lambda|x) = \mathbb{P}(X = x|\lambda),$$

in the continuous case, it is built on a continuous density

$$L(\lambda|x) = f(x|\lambda).$$

The parameter λ does not need to be scalar, inference based on multivariate likelihoods is especially discussed in section 3.

The conventions for notation that we use today go back to Ronald Aylmer Fisher. He started denoting observations with Latin letters and unobserved parameters with Greek letters [11].

Combination of Likelihoods

Due to its definition, likelihoods can be easily combined [16]. When there are independent observations $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_n)$ with assumed probability models $\mathbb{P}_X(\lambda|x)$

and $\mathbb{P}_Y(\lambda|y)$ which are based on the same parameter, the combined likelihood can be specified by the product of the likelihoods referring to the two probability models:

$$L(\lambda|x,y) = L_X(\lambda|x) \cdot L_Y(\lambda|y)$$

2.2 Basic Principles of Likelihood Inference

The likelihood function now can be used to draw conclusions about the unknown parameters that is to say to conduct statistical inference. In the following, some basic principles of likelihood inference are introduced. These deal with the information content in samples and how conclusions can be drawn from these samples when likelihood inference is conducted.

Sufficiency Principle

When x_1 and x_2 depict samples from an experiment X, and $T(x_1) = T(x_2)$ are identical and minimal sufficient statistics from the two samples, the sufficiency principle proves, that the results from statistical inference on the two samples are equal [16]. To follow Pawitan's notation, where evidence of an experiment X with its corresponding outcome x_i is formalized as $Ev(X, x_i)$, i. e.

$$Ev(X, x_1) = Ev(X, x_2).$$

Conditionality Principle

Before introducing the conditionality principle, the term "mixture experiment" needs to be clarified. A mixture experiment is an experiment, which is conducted in two (or more) stages. At a first stage an experiment creates an outcome. This outcome influences which of a number of experiments at a second stage is conducted. The outcome of the second experiment is then the sample of interest.
Now, when x is a sample from a mixture experiment X and x_i is a sample from the actually performed experiment X_i then the conditionality principle proves, that the results from the statistical inference should be conditioned on the experiment that has actually been

performed, i. e.

$$Ev(X, x) = Ev(X_i, x_i).$$

This implies that the structural component of the experiment is not to be considered for statistical inference. To get a better understanding of the conditionality principle, consider the following example

Example by Pawitan

A mixture experiment M is conducted with two measuring devices which perform experiments M_1 and M_2 at a second stage. On a first stage, one of the two devices is arbitrarily chosen $\mathbb{P}(M_1) = \mathbb{P}(M_2) = \frac{1}{2}$. The experiments on the second stage are modeled by normal distributions: $M_1 \sim \mathcal{N}(\mu, 1)$, $M_2 \sim \mathcal{N}(\mu, 4)$.
These two probability models are the models we have to choose when conditioning on the experiment that has actually been performed. When we take into account the whole structure of the experiment, we will obtain a two-component Gaussian mixture model $M \sim \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(\mu, 4)$.
This makes a difference, and this difference can be seen, when we assume that we have performed the experiment and received an observation $x = 3$ from measuring device M_2. When we intend to test the hypothesis H_0: $\mu = 0$ vs. H_1: $\mu > 0$, we can perform an exact test on the given probability models. Hence, we will receive the following p-values:

- $p_M = \mathbb{P}(x > 3|\mu = 0) = \int_3^\infty f(x|\mu)dx = 0.0341$
- $p_{M_1} = \mathbb{P}(x > 3|\mu = 0, M_1) = \int_3^\infty f_{M_1}(x|\mu)dx = .0668$
- $p_{M_2} = \mathbb{P}(x > 3|\mu = 0, M_2) = \int_3^\infty f_{M_2}(x|\mu)dx = .0013$

Now, it is clear that it makes a difference when we condition on the experiment that has actually been performed, because the p-value for the whole experiment and the p-value of the actually conducted experiment are differing from each other. In our example, we know that experiment M_2 has been performed, hence the conditionality principle states we have to use p_{M_2}. The test situation is also shown in figure 1. The p-values are the area under the density curves to the right of the experiments result (dashed line). [1][4][16]

In the following, the weak and the strong likelihood principle are introduced. These principles describe how conclusions can be drawn from likelihood functions that are proportional to each other.

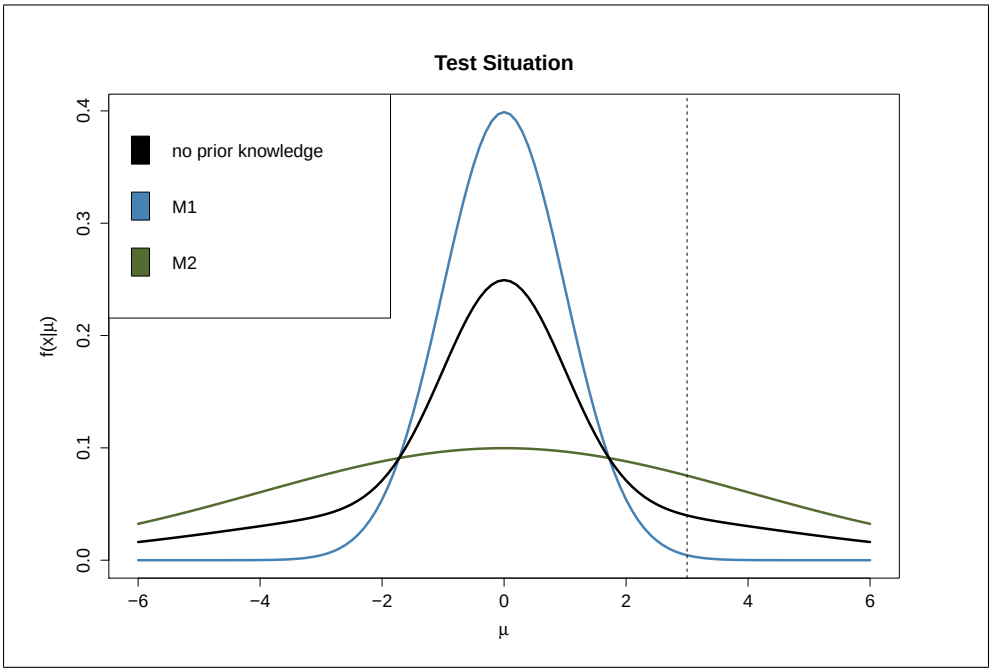

Figure 1: Test Situation

Weak Likelihood Principle

Assuming that we have two samples x_1 and x_2 from an experiment X and that their corresponding likelihoods share a common parameter λ and are proportional to each other

$$L_X(\lambda|x_1) \propto L_X(\lambda|x_2),$$

the weak likelihood principle proves, that the results of statistical inference based on x_1 and x_2 are the same. The weak likelihood principle is equivalent to the sufficiency principle [5][16].

Strong Likelihood Principle

For the strong likelihood principle we assume that there are samples x_1 and x_2 from different experiments X_1 and X_2 as well as their likelihoods share a common parameter λ and are proportional to each other.

$$L_{X_1}(\lambda|x_1) \propto L_{X_2}(\lambda|x_2)$$

Then the strong likelihood principle states that the results of inference based on x_1 and x_2 are the same. Birnbaum's theorem proves that the strong likelihood principle is equivalent to the sufficiency principle and the conditionality principle [2]. However it can be shown that the strong likelihood principle does not always work.

2.3 Limits of the Strong Likelihood Principle

Lindley and Phillips [13] describe a Bernoulli process experiment, in which the likelihood principle is violated. A coin is tossed 12 times, the outcome of the experiment is $x = (x_1, ..., x_{12}) = (0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1)$, where 1 depicts the coin fell on the up side, 0 that it fell on the bottom side. We want to test, if the coin is fair, i. e. $H_0 : \pi = \frac{1}{2}$ vs. $H_1 : \pi > \frac{1}{2}$. However, we do not know, how the number of coin tosses has been determined. In the following, we will discuss two possible assumptions:

Binomial Model
It could be assumed, that the number of coin tosses was predetermined. In this case, we can describe the corresponding probability model by a binomial distribution and hence, we can derive the likelihood:

$$L_b(\pi|x) = \binom{n}{\sum_{i=1}^n x_i} \pi^{\sum_{i=1}^n x_i} (1-\pi)^{n-\sum_{i=1}^n x_i} = \binom{12}{9} \pi^9 (1-\pi)^3$$

In this case, the parameter π stands for the probability of the coin falling on the bottom side.

From the binomial probability model, we can also calculate the p-value for the test in this situation:

$$p_b = \mathbb{P}(\sum_{i=1}^n x_i \geq 9 | \pi = \frac{1}{2}) = \left(\binom{12}{9} + \binom{12}{10} + \binom{12}{11} + \binom{12}{12} \right) \left(\frac{1}{2} \right)^{12} = \frac{299}{4096}$$

Negative Binomial Model
However, it could also be assumed, that the number of coin tosses is determined depending on the outcomes of the experiment. We could assume that the experiment was supposed to stop, after the coin fell three times on the up side. In this case, we can describe the

situation by a negative binomial distribution and again we can easily derive the likelihood function from the corresponding probability model

$$L_{nb}(\pi|x) = \binom{n-1}{\sum_{i=1}^{n} x_i} \pi^{\sum_{i=1}^{n} x_i} (1-\pi)^{n-\sum_{i=1}^{n} x_i} = \binom{11}{9} \pi^9 (1-\pi)^3.$$

The way the negative binomial model is used here, the parameter π is the same as in the previous binomial model and stands for the probability of the coin falling on the bottom side. Thus, both likelihood functions depend on the same parameter and are proportional to each other. However, the results from statistical inference differ, as the p-value for the test in this situation does not match the one in the above example:

$$\begin{aligned} p_{nb} &= \mathbb{P}(\sum_{i=1}^{n} x_i \geq 9 | \pi = \frac{1}{2}) = 1 - \mathbb{P}(\sum_{i=1}^{n} x_i \leq 8 | \pi = \frac{1}{2}) \\ &= 1 - \sum_{i=0}^{8} \binom{i+3-1}{i} \pi^3 (1-\pi)^i = \frac{134}{4096} \neq \frac{299}{4096}. \end{aligned}$$

Hence, in this case, the strong likelihood principle is violated. [13]
At this stage, I would also like to point out that it is possible to construct counterexamples against the conditionality principle, see e.g. Helland [12]. Moreover, I would like to mention that on the one hand, it is possible to base inference on such principles as described above, but on the other hand it is not explicitly necessary to do so and thus, the counterexamples of these principles must be seen within the light of when inference is based on them, one has to keep in mind that they might not be applicable in any situation.

2.4 Methods of Likelihood Inference

After introducing some key principles of likelihood inference, below applicable methods of statistical inference based on the likelihood function will be described. This is to say methods for deriving point estimates (Maximum Likelihood Approach), interval estimates based on the likelihood ratio as well as interval estimates and tests based on asymptotic distribution assumptions (Wald-, score-, likelihood ratio tests and intervals).

Maximum Likelihood

The maximum likelihood method is a technique for deriving point estimates for a parameter of interest based on the likelihood function. The maximum likelihood estimate is the value at which the likelihood function takes its maximum.

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} L(\lambda|x) = \arg\max_{\lambda} \log L(\lambda|x)$$

Because of the monotonic characteristics of the log function, the maximum of the likelihood function and the maximum of the log likelihood function are the same. This can be an advantage for analytically deriving the maximum likelihood estimate from the log likelihood function when this is not possible or very difficult for the likelihood function itself.

Likelihood Ratio

Before describing, how interval estimates can be gained from a likelihood function, the term likelihood ratio shall be explained. The likelihood ratio can be used to compare different values λ_1 and λ_2 for a given likelihood function.

$$\Lambda = \frac{L(\lambda_1|x)}{L(\lambda_2|x)}$$

It can be interpreted as a degree of support for the hypothesis $\lambda = \lambda_1$ against the hypothesis $\lambda = \lambda_2$. When the likelihood ratio $\Lambda > 1$, there will be more support for λ_1 over λ_2, if $\Lambda < 1$ there will be more support for λ_2 over λ_1, respectively.
The likelihood ratio is invariant under bijective transformations of the data $y = f(x)$:

$$\frac{L(\lambda_1|y)}{L(\lambda_2|y)} = \frac{L(\lambda_1|x)}{L(\lambda_2|x)}$$

In the following, we can see how the likelihood ratio can be used to derive interval estimates based on a likelihood function.

Likelihood Interval Estimates

For a given point estimate $\hat{\lambda}$, the likelihood ratio will provide such an interval estimate by defining a set of values λ for which the likelihood ratio of λ and $\hat{\lambda}$ is greater than a predetermined constant c which controls the width of the interval.

$$\{\lambda : \frac{L(\lambda|x)}{L(\hat{\lambda}|x)} > c\}$$

Pawitan mentions that these intervals can only be constructed when λ is scalar. Moreover, it is also not clear how to objectively determine the constant c, as the likelihood itself is not calibrated [16]. However, this interval estimate is a direct likelihood based estimate and no further assumptions on the calibration of the likelihood are made. Still, for practical concerns, it makes sense to constitute further assumptions.

Asymptotic Likelihood Inference

To calibrate the likelihood, we can make assumptions on the asymptotic distributions of some deductions of the likelihood function. These deductions are as already mentioned in the previous example, the likelihood ratio, moreover, the score function as well as the maximum likelihood estimate and asymptotic distributions can be gained by the central limit theorem.

The likelihood ratio is asymptotically χ^2 distributed with one degree of freedom:

$$2\log\frac{L(\hat{\lambda})}{L(\lambda)} \approx J(\hat{\lambda})(\lambda - \hat{\lambda})^2 \overset{as}{\sim} \chi^2(1)$$

This can be used to calculate "highest likelihood" interval estimates

$$\{\lambda : 2\log\frac{L(\hat{\lambda})}{L(\lambda)} \leq \chi^2_{1-\alpha}(1)\},$$

as well as to perform the likelihood ratio test, which is in its two sided case:

$$\phi_\Lambda(x) = \begin{cases} 1, & T_\Lambda > \chi^2_{1-\alpha,1} \\ 0, & \text{sonst} \end{cases}.$$

The score function is asymptotically normally distributed:

$$S(\hat{\lambda}) \overset{as}{\sim} \mathcal{N}(0, J(\hat{\lambda})).$$

The interval estimates based on this distribution assumption are called score intervals

$$\{\lambda : |S(\lambda)| \leq z_{1-\frac{\alpha}{2}} \sqrt{J(\hat{\lambda})}\},$$

the related test is referred to as score test, which is in its two sided case:

$$\phi_S(x) = \begin{cases} 1, & |T_S| \geq z_{1-\frac{\alpha}{2}} \\ 0, & \text{sonst} \end{cases}.$$

Lastly, the maximum likelihood estimate itself is also asymptotically normally distributed

$$\hat{\lambda} \overset{as}{\sim} \mathcal{N}(\hat{\lambda}, J(\hat{\lambda})^{-1}).$$

The interval estimates based on this distribution assumption are called "Wald" intervals

$$[\hat{\lambda} - z_{1-\frac{\alpha}{2}} (J(\hat{\lambda}))^{-\frac{1}{2}}, \hat{\lambda} + z_{1-\frac{\alpha}{2}} (J(\hat{\lambda}))^{-\frac{1}{2}}],$$

the related test is referred to as "Wald" test, which is in its two sided case:

$$\phi_W(x) = \begin{cases} 1, & |T_W| \geq z_{1-\frac{\alpha}{2}} \\ 0, & \text{sonst} \end{cases}$$

In comparison to the interval estimates directly based on the likelihood function without any further assumptions, these three approaches are practically more relevant, as they allow to objectively set a limit for the width of the interval and hence, allow to interpret the intervals in a "probabilistic" sense [16].

2.5 Anomalous Behaviour

Now that some methods for application of the likelihood concept have been shown, in the following two examples, it can be seen that the likelihood concept is not always providing useful inferential results. Below, it is shown how the likelihood function can react anomalously when multiple maxima or singularities occur.

Multiple Maxima

Edwards describes an example, in which a Cauchy model with fixed scale parameter 1 and unknown location parameter λ is used [6].

$$f(x|\lambda) = \frac{1}{\pi(1 + (x - \lambda)^2)}$$

We have two data points $x_1 = -15, x_2 = 15$ and we are interested in a point estimate for the parameter λ. Edwards argues, that as the Cauchy density is symmetrical and we have only two observations, there is only one useful estimate for λ, i. e. the arithmetic mean of x_1 and x_2: $\hat{\lambda} = \frac{x_1+x_2}{2}$. However, when we look at the likelihood function of the model (see figure 2), we can see, that this estimate does not get maximum support at all. As maximum likelihood estimates we receive the two values x_1 and x_2. Thus, as these estimates are strongly differing from the sole useful estimate (the aritmetic mean) the maximum likelihood estimate is not useful in this case. The models for the three estimates can be seen in figure 3. [6]

Singularities

Another example for anomalous behavour is described by Murphy and Bolling [14] [6]. In their article, they describe, that in likelihood functions of single components of mixture distributions which only contain one single observation, singularities can occur. This can be illustrated in the following example. Assume, we have a Gaussian mixture model with density function $f(x|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) = \frac{200}{201}\mathcal{N}(20, 1) + \frac{1}{201}\mathcal{N}(30, 1)$. This means, we have two components that both follow a normal distribution and that are highly unequally mixed ($\pi = \frac{200}{201}$). Now we obtain a sample from this distribution and want to use the expectation-maximization algorithm to fit a Gaussian mixture model to the data. When the sample now suggests that we have only one observation in one of two

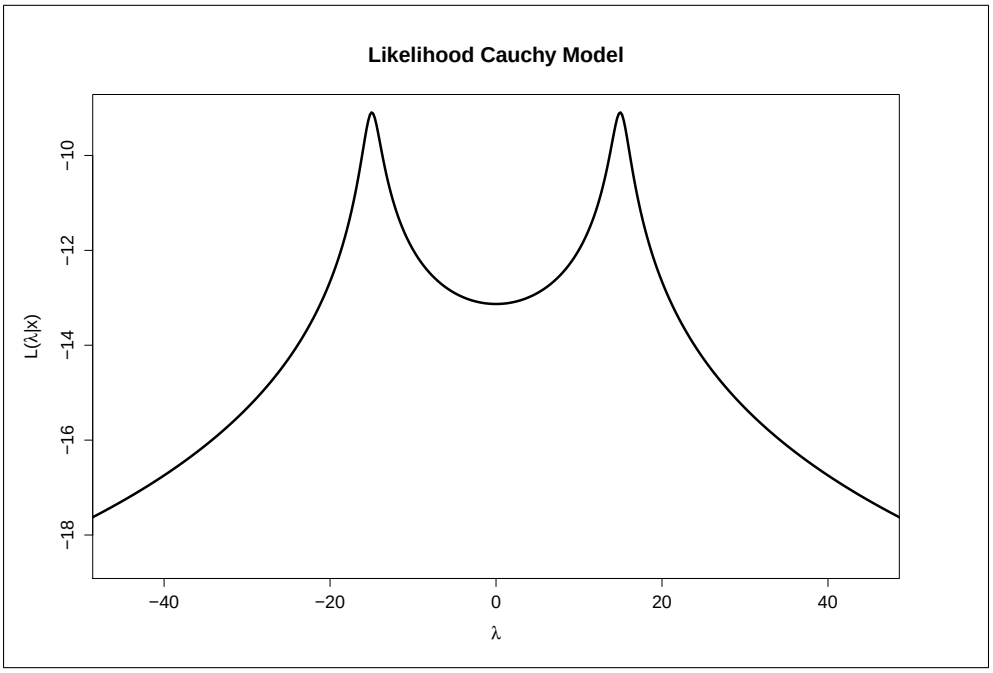

Figure 2: Likelihood of Cauchy Model

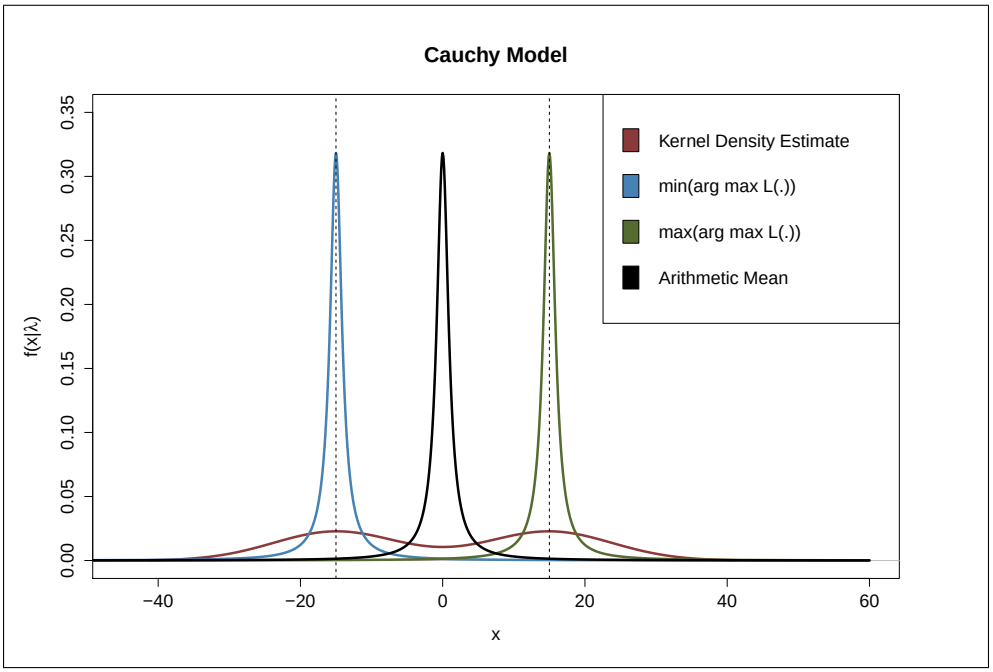

Figure 3: Densities of Different Estimators for Cauchy Model

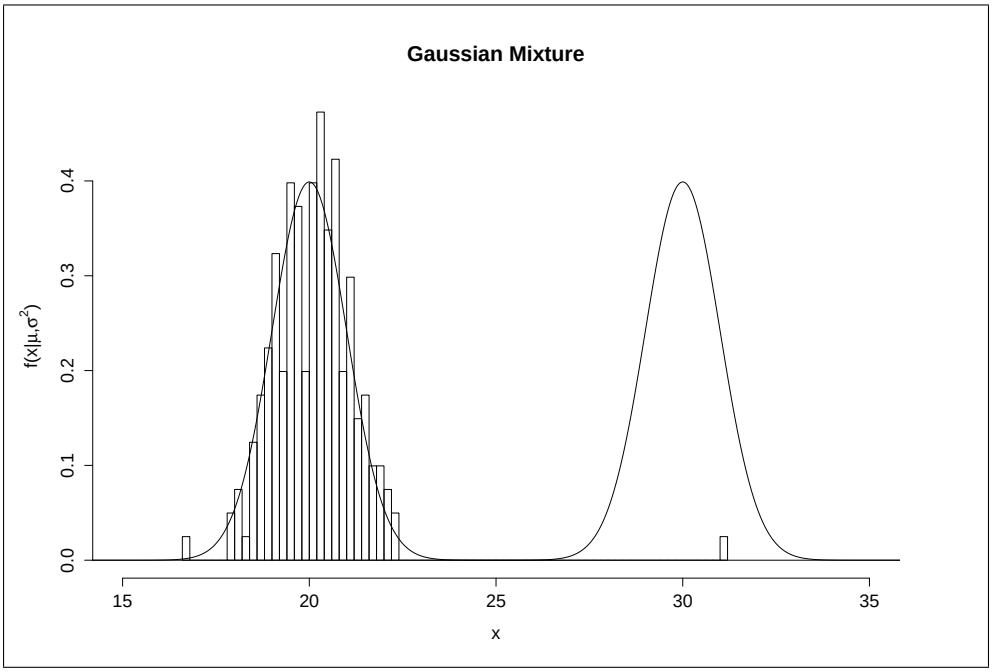

Figure 4: Two Component Gaussian Mixture Model

assumed components (see figure 4), the problem occurs. The likelihood function for the component with the single observation, as can be seen in figure 5 shows a singularity at the point at which μ meets the observed value and $\sigma^2 = 0$. This means that the point estimate $\sigma^2 = 0$ gets infinite support by the likelihood function. However, variance to be zero is an unacceptable assumption. Still this example has practical relevance. When the expectation-maximization algorithm is used, examples like this can happen, especially when a large number of components is assumed for few observations.

3 Multivariate Likelihood, Nuisance Parameters

The parameter λ does not necessarily need to be scalar, it can also be a vector $\lambda = (\lambda_1, ..., \lambda_n)$. An illustration of multivariate likelihoods has already been shown in the previous example. Inference for such multidimensional models can be difficult, when it is conducted for all parameters at once [16]. Sometimes there is also just one (or a few) parameter(s) of interest and inference is only to be relied on these. In this light the following section provides a repertoire of methods for the establishment of likelihood functions for multivariate models in which only the parameters of interest λ_i are captured. Those parameters that are not of interest $\lambda_{-i} = \{\lambda_1, ..., \lambda_n\} \backslash \lambda_i$ are referred to as "nuisance"

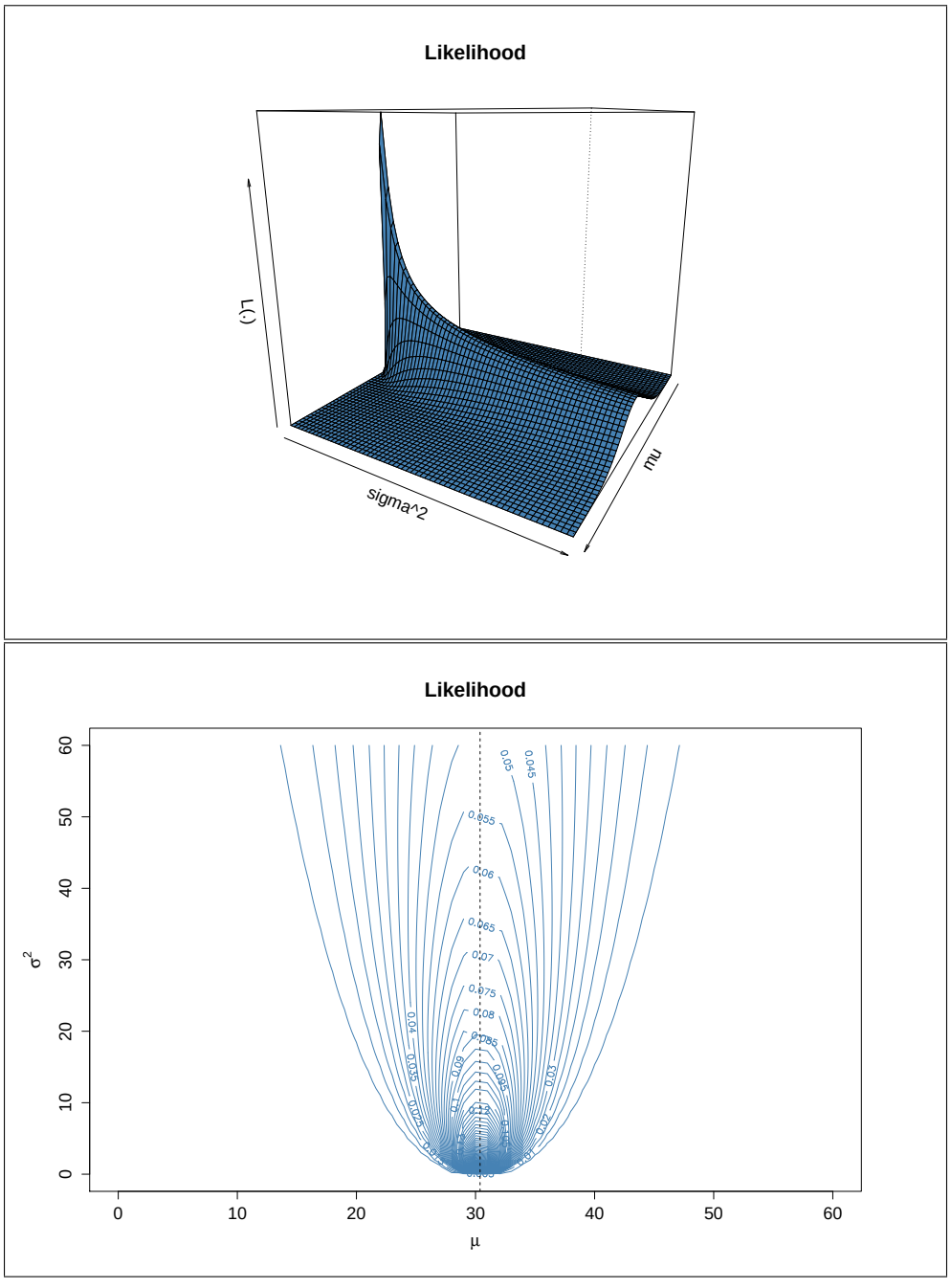

Figure 5: Likelihood for Second Component of Gaussian Mixture Model

parameters.

3.1 Estimated Likelihood

A simplistic approach to achieve this goal is to define the so called estimated likelihood. The estimated likelihood is derived from the likelihood function of all parameters, by replacing the nuisance parameters each by their single maximum likelihood estimate.

$$L_e(\lambda_i) = L(\lambda_i, \hat{\lambda}_{-i}|x) \text{ mit } \hat{\lambda}_{-i} = \{\hat{\lambda}_1, ..., \hat{\lambda}_n\} \backslash \hat{\lambda}_i$$

Now, inference on the parameter of interest can be conducted from the estimated likelihood. However this approach neglects the fact that the nuisance parameters are unknown and hence, this procedure can be problematic when applied [16].

3.2 Profile Likelihood

The profile likelihood concept offers a different approach. It does not replace the nuisance parameters by single maximum likelihood estimates, instead it maximizes the likelihood function for all nuisance parameters at once.

$$L_p(\lambda_i) = \max_{\lambda_{-i}} L(\lambda_1, ..., \lambda_n|x) \text{ mit } \lambda_{-i} = \{\lambda_1, ..., \lambda_n\} \backslash \lambda_i$$

This approach involves multivariate optimization. This can be either done analytically or by using numerical methods, e.g. by using the Newton-Raphson method, the Quasi-Newton method or the Simplex method by Nelder and Mead.
Inference on the parameter of interest can then be based on the profile likelihood. In some cases, estimated likelihood and profile likelihood are equal.

3.3 Limits of the Profile Likelihood Concept

However, as the profile likelihood concept can be considered better than the estimated likelihood approach, there is also a drawback which shall be discussed here. The profile likelihood can be severely biased in some cases, as the following example by Neyman and Scott shows [15] [16]. They describe a stratified dataset. There are N levels. For each

level, we have two observations and for each level we assume a normal distribution model with specific location parameter μ_i, but the same variance in all levels σ^2. This means that in total we have $N+1$ parameters: $\theta = \{\mu_1, ..., \mu_N, \sigma^2\}$. Table 3.3 shows a possible, simulated outcome of this model for given parameters μ_i and true $\sigma^2 = 1$. Now as we take the given location parameters as unknown, we can derive the profile likelihood of σ^2 and by knowing the location parameters μ_i we can also calculate the true likelihood of σ^2 by simply feeding the true μ_i values into the Likelihood function of all parameters, which is

$$L(\theta|x_i) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{2}(x_{ij}-\mu_i)^2\right),$$

so that there is only the unknown σ^2 left. Hence, we can compare the profile likelihood and the true likelihood for σ^2. To better see the differences between those two, they are both scaled in a way that their maximum values are 1. The result can be seen in figure 6. The underlying source code for the maximization of the likelihood function concerning the nuisance parameters in order to derive the profile likelihood can be found in the appendix. In figure 6 we can see that the true likelihood has its maximum at the true σ^2. In fact, this does not necessarily need to be the case as the data can also differ more strongly and give maximum support to an other value than $\sigma^2 = 1$. Still we can see, that the profile likelihood and the true likelihood for σ^2 are differing from each other.

However, to see that the profile likelihood is biased, we have to analytically derive the profile likelihood. In this example, profile likelihood and estimated likelihood are the same, i. e. the single maximum likelihood estimates $\hat{\mu}_i = \bar{x} = \frac{x_{i1}+x_{i2}}{2}$ for the nuisance parameters μ_i will be the same as the global maximization of the overall likelihood function concerning the nuisance parameters. This example is chosen in a way that it reflects an analysis of variance (ANOVA) model. In this model the part within the exponent of the likelihood function in which the single estimates appear can be described as the residual sum of squares $\sum_{i=1}^{N}\sum_{j=1}^{2}(x_{ij}-\hat{\mu}_i)^2 = \sum_{i=1}^{N}\sum_{j=1}^{2}(x_{ij}-\bar{x}_i)^2 = RSS$. Hence, when we want to derive the maximum likelihood estimate of σ^2 based on its profile likelihood

$$L_p(\sigma^2) = \max_{\mu_1,...,\mu_N} L(\theta|x_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}RSS\right),$$

which involves deriving its log likelihood function

$$l_p(\sigma^2) = -N\log(2\pi\sigma^2) - \frac{1}{2\sigma^2})RSS,$$

i	μ_i	y_{i1}	y_{i2}	i	μ_i	y_{i1}	y_{i2}
1	-3.97	-4.50	-3.65	21	-4.50	-2.65	-4.55
2	0.99	0.98	2.58	22	10.70	10.44	8.53
3	5.01	6.34	4.89	23	-5.83	-5.72	-4.52
4	6.44	6.53	6.42	24	-0.24	-1.13	-1.28
5	4.53	3.70	5.23	25	2.08	1.54	1.75
6	2.47	2.27	2.31	26	8.60	10.27	7.55
7	3.00	3.22	3.30	27	-3.92	-3.84	-3.22
8	-7.90	-6.53	-9.29	28	-6.52	-5.01	-6.52
9	5.00	3.36	6.43	29	-2.26	-0.83	-3.37
10	10.94	11.04	11.25	30	-8.87	-8.30	-7.53
11	-6.05	-5.16	-5.15	31	-0.25	-0.18	0.94
12	-2.94	-1.43	-2.43	32	-5.19	-7.30	-4.52
13	5.28	4.94	6.65	33	3.11	2.92	2.92
14	-1.58	-0.76	-0.16	34	0.64	1.42	1.20
15	-0.27	1.92	-1.17	35	-8.70	-8.72	-8.96
16	1.65	-0.04	1.18	36	-2.87	-3.03	-4.96
17	3.32	2.96	4.06	37	4.56	5.01	3.86
18	4.39	4.56	6.04	38	6.17	6.95	5.91
19	1.01	1.05	-0.02	39	3.31	3.33	3.71
20	11.37	10.74	11.95	40	-4.31	-2.34	-3.35

Table 1: Observations and True Location Parameters

calculating its derivative for σ^2 to receive the score function and solving the score equation

$$S_p(\sigma^2) = -\frac{N}{\sigma^2} + \frac{1}{2(\sigma^2)^2} RSS \stackrel{!}{=} 0,$$

we will receive

$$\hat{\sigma^2}_{ML} = \frac{1}{2N} RSS.$$

as maximum likelihood estimate.

In the ANOVA model, we know the distribution of the ratio of the residual sum of squares and the true variance parameter

$$\frac{RSS}{\sigma^2} \sim \chi^2_N$$

can be described by a χ^2 distribution [16] with N degrees of freedom. Hence, we can calculate the expectation and variance of the maximum likelihood estimate $\hat{\sigma^2}_{ML}$:

$$\mathbb{E}(\hat{\sigma^2}_{ML}) = \mathbb{E}\left(\frac{RSS}{2N}\right) = \frac{\sigma^2}{2N}\mathbb{E}\left(\frac{RSS}{\sigma^2}\right) = \frac{\sigma^2}{2N} \cdot N = \frac{\sigma^2}{2}$$

$$\mathbb{V}(\hat{\sigma^2}_{ML}) = \mathbb{V}\left(\frac{RSS}{2N}\right) = \frac{\sigma^4}{4N^2}\mathbb{V}\left(\frac{RSS}{\sigma^2}\right) = \frac{\sigma^4}{4N^2} \cdot 2N = \frac{\sigma^4}{2N}$$

The expectation of the maximum likelihood estimate is not equal to σ^2, hence biased, and the bias does also not diminish asymptotically. Moreover, considering both, bias and variance, we can see the maximum likelihood estimate is also not MSE consistent.

$$\hat{\sigma^2}_{ML} \xrightarrow{p} \frac{\sigma^2}{2}$$

At this stage, it should be pointed out that bias is a general problem of the likelihood approach and not only limited to this example. However, in this example, the bias is severe. Moreover, one should bear in mind that Neyman and Scott, although describing the problem as introduced here, did not use the term "profile likelihood" in their publication as this term did not exist in 1948 [*]. In a paragraph on the history of profile likelihood Sprott [18] argues that almost 20 years later, in 1964 Box and Cox used the profile likelihood under the name "maximized likelihood" because the term "profile likelihood" was also not known then.

3.4 Marginal and Conditional Likelihood

Two other methods for deriving likelihoods for parameters of interest in the context of multivariate likelihoods are the marginal likelihood concept as well as the conditional likelihood approach. Both are based on the idea of transforming the observation vector x into two vectors $x \to (a, b)$.

Marginal Likelihood

For the marginal likelihood this transformation is that the density function for the data in the model based on all parameters can be divided in a way that the marginal density of the transformed vector a is only dependent on the parameter of interest λ_i.

$$L(\lambda_i, \lambda_{-i}|a, b) = f(a, b|\lambda_i, \lambda_{-i}) = f(a|\lambda_i) f(b|a, \lambda_i, \lambda_{-i}) = \boldsymbol{L_m(\lambda_i)} L_\epsilon(\lambda_i, \lambda_{-i})$$

The marginal likelihood is then defined by this marginal density.

Conditional Likelihood

The conditional likelihood approach is similar. The idea here is that not the marginal but the conditional density of the data vector $a|b$ is only dependant on the parameter of

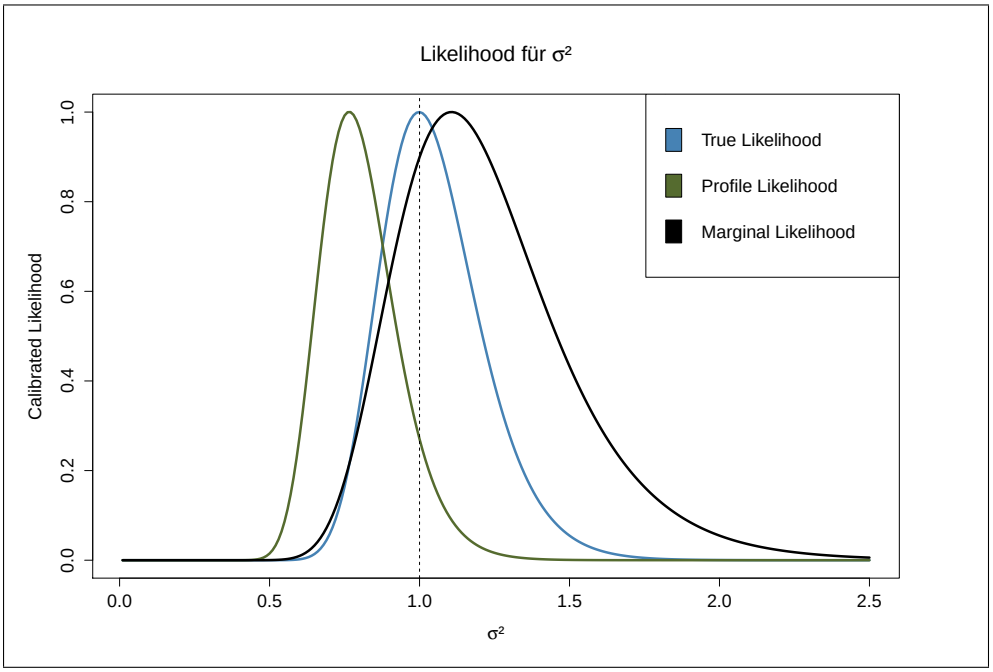

Figure 6: Marginal Likelihood

interest λ_i.

$$L(\lambda_i, \lambda_{-i}|a, b) = f(a|b, \lambda_i)f(b|\lambda_i, \lambda_{-i}) = \boldsymbol{L_c(\lambda_i)}L_\epsilon(\lambda_i, \lambda_{-i})$$

The conditional likelihood is then defined by this conditional density. Conditional Likelihood generally exists when all parameters, λ_i as well as the nuisance parameters are the natural parameters of an exponential family model [16].
Nevertheless, marginal and conditional likelihoods are not always availably or are difficult to be obtained.

Marginal Likelihood - Example by Neyman and Scott

According to Pawitan, the marginal likelihood can be a useful approach for the previous example concerning the biased profile likelihood. When we transform the data $x \to (a, b)$ in the following manner

$$a_i = \frac{x_{i1} - x_{i2}}{\sqrt{2}}, b_i = \frac{x_{i1} + x_{i2}}{\sqrt{2}},$$

we will receive an unbiased maximum likelihood estimate from the marginal likelihood

$$L_m(\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} a_i^2\right).$$

The marginal likelihood function can be seen in figure 6. When it is mentioned that the marginal likelihood is unbiased, however this does not mean that the maximum of the marginal likelihood is the same as the maximum of the true likelihood function.

The fact that the maximum likelihood estimate of the marginal likelihood function is unbiased can again be seen when we use the information that within an ANOVA model, the ratio of the residual sum of squares and the true variance is χ^2 distributed with N degrees of freedom. Because then

$$\mathbb{E}(\hat{\sigma^2}_{ML}) = \mathbb{E}(\frac{1}{N}\sum_{i=1}^{N} a_i^2) = \mathbb{E}(\frac{RSS}{N}) = \sigma^2$$

the expectation of the maximum likelihood estimate is the true parameter σ^2 and hence it is shown that the maximum likelihood estimate is unbiased.

4 Comparison to Other Uncertainty Concepts

After the foundations of the likelihood concepts have been introduced and methods of inference have been discussed in the single and multi-parameter context, in the following section, the historical use and development is depicted. Furthermore the likelihood concept is compared to both, the Bayesian and frequentist approaches. Lastly it is shown how the likelihood concept is discussed in scientific literature.

4.1 Historical Development

The likelihood concept, as we use it today goes back to Ronald Aylmer Fisher (1890-1962). He introduced the term “likelihood” in 1921 and published the maximum likelihood method in 1922. However, the idea of the maximum likelihood concept goes further back in history. It was already used in the 18th century and was independently developed

by Johann Heinrich Lambert (1728-1777) and Daniel Bernoulli (1700-1782) [11]. Laplace (1749-1827) is also known for developing the concept of "inverse probability", which is referring to the likelihood function. As already mentioned at the beginning, "inverse probability" means, an unknown parameter is looked at given an observed sample $\lambda|x$ which can be viewed as "inverse" to the probability of an observed sample given an unobserved parameter $x|\lambda$. Carl Friedrich Gauß (1777-1855) used the approaches by Lambert, Bernoulli and Laplace to perform parameter estimates for his method of least squares.

4.2 Bayesian Approach

Now, it shall be discussed how the likelihood concept can be compared to the Bayesian approach. The Bayesian view on probability is that it is a subjective degree of belief on a parameter or a future observation. In a Bayesian model there is a prior assumption on the distribution of the parameter.

$$f(\lambda|x) \propto f(x|\lambda)f(\lambda)$$

In Bayesian statistics, inference is carried out from the posterior distribution. The posterior distribution equals the likelihood function when we assume a constant prior distribution for the unknown parameter, i. e. $p(\lambda) \equiv 1$ and calibrate the function in order to integrate to 1. In this case, we assume that the prior distribution does not contain information about the parameter λ. If there is prior knowledge about the parameter available, the likelihood concept can not deal with it.

Here, it should be mentioned that although Pawitan describes that assuming a uniform distribution is equivalent to having no prior knowledge is not true, because the assumption of a constant prior is also an implication of knowledge concerning the plain structure of the prior distribution [*].

4.3 Frequentist Approach

Frequentists define probability differently. They see it as the limit of a random sequence of relative frequencies. For true frequentists likelihood inference can be pointless in some cases when it is referring to a parameter. When we consider a 95% confidence interval for

a parameter $3.3 \leq \lambda \leq 5.2$, frequentists could not make a conclusion on the parameter based on this interval, because the 95% confidence level is referring to the confidence interval procedure and not to the parameter from their perspective [16].

Repeated Sampling Principle

One fundamental property of the frequentist approach is expressed in the "repeated sampling principle". This principle states that inferential methods should be based on their behaviour in hypothetical repetitions under the same conditions [16]. This, however can be in conflict with the likelihood concept, as the following example shows.

Example by Fraser et al.

Fraser et al. [8] describe a situation in which the repeated sampling principle is violated. Assume we conduct an experiment with a given density function $f(x|\lambda)$, and in which the sample space is a triple dependent on an unobserved parameter λ.

$$f(x|\lambda) = \begin{cases} \frac{1}{3}, & x = 1, 2, 3 & \lambda = 1 \\ \frac{1}{3}, & x = \frac{\lambda}{2}, 2\lambda, 2\lambda + 1, & \lambda \text{ gerade} \\ \frac{1}{3}, & x = \frac{\lambda - 1}{2}, 2\lambda, 2\lambda - 1, & \lambda \text{ ungerade} \end{cases}$$

$\lambda \in \mathbb{N}$ and $x \in \mathbb{N}$ are scalar. The density function is constant. To get a better idea of this density function, consider table 2.

		λ							
		1	2	3	4	5	6	7	8
x	1	1	1	1	0	0	0	0	0
	2	1	0	0	1	1	0	0	0
	3	1	0	0	0	0	1	1	0
	4	0	1	0	0	0	0	0	1
	5	0	1	0	0	0	0	0	0
	6	0	0	1	0	0	0	0	0
	7	0	0	1	0	0	0	0	0
	8	0	0	0	1	0	0	0	0

Table 2: Incidence Matrix (see Fraser et al. [8])

As we know the probability density function, we can also define the corresponding likelihood function, which is also constant.

$$L(\lambda|x) = \begin{cases} \frac{1}{3}, & \lambda = 1,2,3 & x = 1 \\ \frac{1}{3}, & \lambda = \frac{x}{2}, 2x, 2x+1, & x \text{ gerade} \\ \frac{1}{3}, & \lambda = \frac{x-1}{2}, 2x, 2x-1, & x \text{ ungerade} \end{cases}$$

When the likelihood function is always constant, this means that every possible value for λ is equally suitable as an estimate for λ, because the likelihood takes its maximum at each value.
However, we can also use a different approach for deriving an estimate for the parameter λ. Due to the way the underlying density function is defined, there are always three different possibilities for a point estimate for λ: $\hat{\lambda}_1 = \lambda_{(1)}, \hat{\lambda}_2 = \lambda_{(2)}, \hat{\lambda}_3 = \lambda_{(3)}$, i. e. we can depict these as the smallest, the middle and the largest value, λ can take for given observations x. With this idea in mind we can now calculate the probabilities that these estimators are correctly determining the unknown parameter:

$$\mathbb{P}(\hat{\lambda}_1 = \lambda) = \begin{cases} 1, & \lambda = 1 \\ \frac{2}{3}, & \lambda > 1 \end{cases}, \mathbb{P}(\hat{\lambda}_2 = \lambda) = \begin{cases} \frac{1}{3}, & \lambda \in 2\mathbb{N} \\ 0, & \lambda \in 2\mathbb{N}-1 \end{cases}, \mathbb{P}(\hat{\lambda}_3 = \lambda) = \begin{cases} \frac{1}{3}, & \lambda \in 2\mathbb{N}+1 \\ 0, & \lambda \in 2\mathbb{N} \end{cases}$$

The calculation of the probabilities can be better understood from the incidence matrix (table 2). From these probabilities there is clear evidence, that $\hat{\lambda}_1$ is a better estimate than $\hat{\lambda}_2$ and $\hat{\lambda}_3$, because its probability of correctly estimating the true parameter is always larger than for the other two estimators:

$$\mathbb{P}(\hat{\lambda}_1 = \lambda) \geq \frac{2}{3} > \frac{1}{3} \geq \mathbb{P}(\hat{\lambda}_2 = \lambda) = \mathbb{P}(\hat{\lambda}_3 = \lambda).$$

This example shows that the likelihood does not account for long term repetitions under the same conditions and hence the repeated sampling principle is violated. Pawitan argues that the probability of correctly estimating the parameter of interest is not a property of information which is stored in the data. Therefore, this information can not be considered in the likelihood function. A similar example is also discussed by Goldstein and Howard

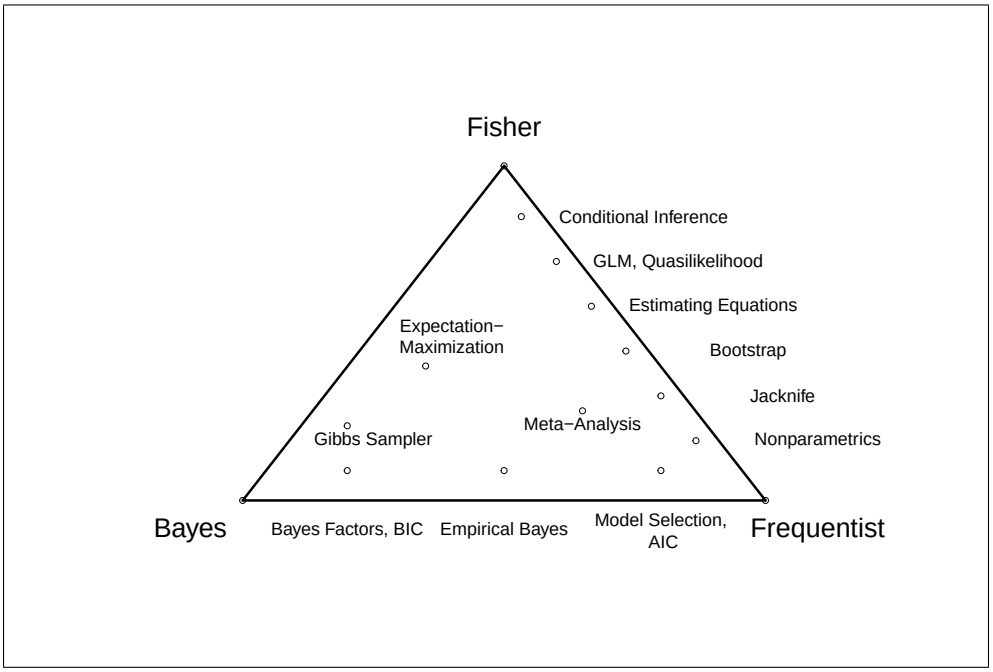

Figure 7: see Efron [7]

[10] [16].

4.4 Discussion About Likelihood Concept

In this last section, I would like to describe how the likelihood concept is viewed in the scientific literature. Pawitan mentions the likelihood approach as a Bayesian-frequentist compromise [16]. Ronald Aylmer Fisher himself explains his point of view when likelihood inference should be used as when there is no possibility of making exact probability statements. He also states, that if sample size is large enough, inference based on asymptotic distributions can also be conducted [9]. Within this light, Pawitan describes Fishers attitude towards probability that he is neither in favor of the strict assumption of a prior distribution in the Bayesian sense, nor a supporter of the strict frequentist definition of probability as a limit of a random sequence of relative frequencies. Following Fishers idea, as described above, his attitude is also not towards likelihood being the sole carrier of uncertainty in statistical inference [16]. Nevertheless, there are also sources available that favour this idea (see e.g. Royall [17]). As a final idea I would like to quote an illustration from Efron [7]. He tries to visualize the tendency of common statistical inference procedures towards a Bayesian or frequentist approach or an approach based on Fisher's ideas. His conclusion can be seen in figure 7 and is an interesting idea, which should be

considered as open for discussion.

In this example, Efron does not only refer to the likelihood approach itself, as he uses the term "Fisher" to describe this approach [*]. By the word "Fisher" he also aims to consider Fisher's idea of "fiducial inference". The fiducial argument was supposed to be an alternative to the Bayesian argument, which was neglected by Fisher because of the mostly arbitrary and non-objective choice of a prior distribution. Unlike in Bayesian inference, in fiducial inference no prior distribution is assumed. However, the fiducial inference approach lacks of an exact definition and has some serious drawbacks as there are problems with multivariate parameter estimation for instance [3][*].

5 Summary

The likelihood concept offers a variety of inferential methods. It has been shown how its basic principals of statistical inference work. Furthermore a broad repertoire of methods has been discussed with a special emphasis on multivariate likelihoods and the dealing with nuisance parameters. A variety of examples proved, that the likelihood concept can not be applied in any situation and that there are shortcomings one should bear in mind when using such inferential methods. Lastly it was shown how the likelihood concept can be compared to Bayesian and frequentist ideas and that there are parallels to these approaches in both cases. The biggest advantage explaining the various applications of likelihood might be its simplicity as Efron states that Fisher's work has "unique quality of [...] mathematical synthesis combined with the utmost practicality" [7].

6 Comparison with other Concepts of Uncertainty

This last section is referring to the seminar "Probability and other Concepts of Uncertainty" again. It aims to compare the likelihood approach to other concepts of uncertainty which have been presented in the seminar. The first talk was about "subjective probability" and this is closely related to Bayesian inference. As described in this work it is possible to show relations between the Bayesian and the likelihood inference concept and hence to subjective probability interpretations. Another talk focussed on "fuzzy sets". In

this talk it was pointed out that likelihood functions can be identified as a special case of fuzzy sets and hence a relation to the likelihood concept can be drawn. This is different with the topic "interval probabilities". The likelihood concept as introduced here cannot deal with such generalization of probability. Moreover, this points out a major difference between likelihood and more sophisticated uncertainty concepts: the likelihood concept is highly applicable, but some other approaches are too complex or deal with data, which only rarely exists (e.g. fuzzy data), so that these concepts are of minor practical relevance [*].

7 References

[1] J. O. Berger and R. L. Wolpert. *The Likelihood Principle*. IMS Lecture Notes Series, Volume 6, 1988.

[2] Alan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Society*, 6:269–305, 1962.

[3] J. F. Box and A. W. F. Edwards. Fisher, Ronald Aylmer. *Encyclopedia of Biostatistics*, 2005.

[4] D. R. Cox. Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics*, 29:357–372, 1958.

[5] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall, London, 1979.

[6] A. W. F. Edwards. *Likelihood*. Cambridge University Press, London, 1972.

[7] B. Efron. Ronald Aylmer Fisher in the 21st century. *Statistical Science*, 13(2):95–122, 1998.

[8] G. A. S. Fraser et al. Marginalization, Likelihood and Structured Models. *Journal of Multivariate Analysis*, 57(298):209–217, 1984.

[9] R. A. Fisher. *Statistical Methods and Scietific Inference*. Hafner Press, 1973.

[10] M. Goldstein and J. V. Howard. A Likelihood Paradox. *Journal of the Royal Statistical Society B*, 51(3):619–628, 1991.

[11] A. Hald. On the history of Maximum Likelihood in Relation to Inverse Probability and Least Squares. *Statistical Science*, 14:214–222, 1999.

[12] I. S. Helland. Simple Counterexamples against the conditionality principle. *The American Statistician*, 49:351–356, 1995.

[13] D. V. Lindley and L. D. Phillips. Inference for a Bernoulli Process (A Bayesian View). *The American Statistician*, 30(3):112–119, 1976.

[14] E. A. Murphy and D. R. Bolling. Testing of single locus hypotheses where there is incomplete separation of the phenotypes. *American Journal of Human Genetics*, 19:322–334, 1967.

[15] J. Neyman and E. L. Scott. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16(1):1–32, 1948.

[16] Y. Pawitan. *In All Likelihood*. Oxford University Press, New York, 2001.

[17] R. M. Royall. *Statistical Evidence: A Likelihod Paradigm*. Chapman & Hall, 1997.

[18] D. A. Sprott. *Statistical Inference in Science*. Springer, 2000.

8 Appendix

R-Code for Bias of Profile Likelihood Example

```
set.seed(41)
N<-40
mu<-rnorm(N,0,5)
sigma<-1
# Generation of dataset
y<-matrix(0,N,2)
for(i in 1:N){
  y[i,1]<-rnorm(1,mu[i],sigma)
  y[i,2]<-rnorm(1,mu[i],sigma)
}
### True Likelihood for sigma^2
likelihood_true<-function(sigma_squared, mu){
  result<-NULL
  for(j in 1:length(sigma_squared)){
    result_step<-1
    for(i in 1:N){
      result_step<-result_step*1/(2*pi*sigma_squared[j])*
        exp(-1/(2*sigma_squared[j])*((y[i,1]-mu[i])^2+(y[i,2]-mu[i])^2))
    }
    result<-c(result, result_step)
  }
  return(result)
```

```
}
vec<-seq(0.01,2.5,.01)
likelihood_true_vec<-likelihood_true(vec,mu)
likelihood_true_max<-max(likelihood_true(vec,mu))
likelihood_true_norm<-likelihood_true_vec/likelihood_true_max
### Profile Likelihood for sigma^2
likelihood_profile <- function(sigma_squared){
  result<-NULL
  for(i in 1:length(sigma_squared)){
    func_optim <- function(nuisance, method="Nelder-Mead"){
      -likelihood_true(sigma_squared[i], mu=nuisance)
    }
    nuisance <- optim(mu, fn = func_optim)
    mu <- nuisance$par
    result<-c(result,likelihood_true(sigma_squared[i], mu))
  }
  return(result)
}
likelihood_profile_vec<-likelihood_profile(vec)
likelihood_profile_max<-max(likelihood_profile_vec)
likelihood_profile_norm<-likelihood_profile_vec/likelihood_profile_max
plot(vec,likelihood_true_norm,t="l",col="steelblue",lwd=3,
    xlab=expression(paste(sigma,"^2")),ylab="Standardized Likelihood",
    main=expression(paste("Likelihood for ",sigma,"^2")))
lines(vec,likelihood_profile_norm, lwd=3, col="darkolivegreen")
abline(v=1,lty=2)
```

R-Code for Marginal Likelihood Example

```
# Transformation of Data
a<-(y[,1]-y[,2])/sqrt(2)
b<-(y[,1]+y[,2])/sqrt(2)
# Marginal Likelihood Function
likelihood_marginal<-function(sigma_squared){
  result<-NULL
  for(j in 1:length(sigma_squared)){
    prod<-1
    for(i in 1:length(a)){
      prod<-prod*dnorm(a[i],0,sqrt(sigma_squared[j]))
    }
    result<-c(result,prod)
  }
  return(result)
}
likelihood_marginal_vec<-likelihood_marginal(vec)
likelihood_marginal_max<-max(likelihood_marginal_vec)
likelihood_marginal_norm<-likelihood_marginal_vec/likelihood_marginal_max
lines(vec,likelihood_marginal_norm,lwd=3)
```

R-Code for Cauchy-Model

```
x<-c(-15,15)
lh<-function(lambda,x){
  result<-NULL
  for(j in 1:length(lambda)){
```

```
    result_new<-0
    for(i in 1:length(x)){
      result_new<-result_new-(log(pi)+log(1+(x[i]-lambda[j])^2))
    }
  result<-c(result,result_new)
  }
  return(result)
}
lambda<-seq(4*x[1],4*x[2],.2)
plot(lambda,lh(lambda,x),t="l",lwd=3,main="Likelihood Cauchy-Modell",
     xlab=expression(lambda),ylab=expression(paste("L(",lambda,"|x)")),
     xlim=c(3*x[1],3*x[2]))
## Cauchy distribution for -50, 50, Kernel Density Estimator
dcauchy<-function(x,theta){
  result<-1/(pi*(1+(x-theta)^2))
  return(result)
}
vec<-seq(4*x[1],4*x[2],.1)
plot(density(x),ylim=c(0,.35),xlim=c(3*x[1],4*x[2]),lwd=3,
     col="indianred4",main="Cauchy-Modell",xlab="x",
     ylab=expression(paste("f(x|",lambda,")")))
lines(vec,dcauchy(vec,x[1]),lwd=3,t="l",col="steelblue")
lines(vec,dcauchy(vec,x[2]),lwd=3,col="darkolivegreen")
lines(vec,(dcauchy(vec,0)),lwd=3)
abline(v=x[1],lty=2)
abline(v=x[2],lty=2)
```

R-Code for Mixture Model

```
par(mar=c(5,5,5,5))
x1<-rnorm(200,20,1)
x2<-rnorm(1,30,1)
hist(c(x1,x2),breaks=100, freq=F, xlim=c(15,35),main="Mischverteilung",
     xlab="x", ylab=expression(paste("f(x|",mu,",",sigma^2,")")))
vec<-seq(0,60,.01)
lines(vec,dnorm(vec,20,1))
lines(vec,dnorm(vec,30,1))
lh<-function(mu,sigma_squared){
    result<-dnorm(x2,mu,sqrt(sigma_squared))
  return(result)
}
vec1<-seq(0,60,1)
vec2<-vec1
z<-outer(vec1,vec2,lh)
contour(vec1,vec2,z,60,main="Likelihood",xlab=expression(mu),
        ylab=expression(sigma^2),col="steelblue")
abline(v=x2,lty=2)
persp(vec1,vec2,z,col="steelblue",theta=120,xlab="mu",ylab="sigma^2",
      zlab="L(.)",main="Likelihood")
```