Scientific Research Publishing

Open Journal of Statistics, 2016, 6, 1037-1049
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Some Likelihood Based Properties in Large Samples: Utility and Risk Aversion, Second Order Prior Selection and Posterior Density Stability

Michael Brimacombe

Department of Biostatistics KUMC, Kansas City, USA
Email: mbrimacombe@kumc.edu

How to cite this paper: Brimacombe, M. (2016) Some Likelihood Based Properties in Large Samples: Utility and Risk Aversion, Second Order Prior Selection and Posterior Density Stability. *Open Journal of Statistics*, **6**, 1037-1049.
http://dx.doi.org/10.4236/ojs.2016.66084

Received: September 8, 2016
Accepted: November 28, 2016
Published: December 2, 2016

Abstract

The likelihood function plays a central role in statistical analysis in relation to information, from both frequentist and Bayesian perspectives. In large samples several new properties of the likelihood in relation to information are developed here. The Arrow-Pratt absolute risk aversion measure is shown to be related to the Cramer-Rao Information bound. The derivative of the log-likelihood function is seen to provide a measure of information related stability for the Bayesian posterior density. As well, information similar prior densities can be defined reflecting the central role of likelihood in the Bayes learning paradigm.

Keywords

Arrow-Pratt Theorem, Expected Utility, Information Similar Priors, Likelihood Function, Prior Stability, Score Function, Risk Aversion

1. Introduction

Research Background

The importance of the likelihood function to statistical modeling and parametric statistical inference is well known, from both frequentist and Bayesian perspectives. From the frequentist perspective the likelihood function yields minimal sufficient statistics, if they exist, as well as providing a tool for the generating of pivotal quantities and measures of information on which to base estimation and hypothesis testing procedures [1].

For researchers employing a Bayesian perspective the likelihood function is modulated into a probability distribution directly on the parameter space through the use of a prior density and Bayes theorem [2]. The Bayesian context preserves the whole of the

likelihood function and allows for the use of probability calculus on the parameter space Ω itself. This usually takes the form of averaging out unwanted parameters in order to obtain marginal distributions for parameters of interest.

Current Research

Research into the properties of the likelihood function has often focused on the properties of the maximum likelihood estimator, and likelihood ratio based testing of hypotheses [3]. A review can be found in [4]. As well, recent work has examined likelihood based properties in relation to saddlepoint approximation based limit theorem results [5]. The Cramer Rao bound or Fisher information continues to be of interest across a wide set of applied fields [6], providing a measure of overall accuracy in the modeling process. Information theoretic measures based on likelihood, such as the AIC measure [7] are commonly applied to assess relative improvement in model predictive properties.

From a Bayesian perspective much recent work has focused on the application of Markov Chain Monte Carlo (MCMC) based approximation and methodology [8] [9]. The algorithms that have been developed in these settings have greatly widened the areas of application for the Bayesian interpretation of likelihood [10].

Prior density selection has often focused on robustness issues [11] where the sensitivity of the posterior density to the selected prior is of interest. Some focus has also been given to choose priors in order to match frequentist and Bayesian inference in terms of choosing priors that match p-values and posterior probabilities, so-called first order matching [12]. Here a focus is placed on large samples and the broader concept of information.

The application of utility theory in a Bayesian context reflects several possible definitions and approaches [2] and some of these are discussed below. This however has been viewed independently of the likelihood concept with utility functions typically assumed in addition to the assumed prior. Here a learning perspective regarding how information is collected and processed through the parametric model in large samples is considered with the likelihood function and the related score function playing key roles in the interpretation of the posterior density from several perspectives.

Research Approach and Strategy

In this paper several large sample properties of the likelihood and their connections to ideas in economics are examined. The derivative of the log-likelihood function is shown to define an elasticity based measures of stability for the posterior density. It is then argued that the log-likelihood function can itself serve as a utility function in large samples, connecting probability based preferences and expected utility optimization with statistical optimization, especially in relation to the consumption of information.

The Bayesian perspective provides the context for this approach, yielding a probability-likelihood pair that allows us to relate expected utility maximization with optimal statistical inference and large sample properties of the likelihood function. From this perspective the well-known Arrow-Pratt risk aversion theorem is shown to be a function of the standardized score statistic and Cramer-Rao Information bound.

2. Fundamental Principles

The likelihood function can be written;

$$L(\theta \mid data) = k \cdot \prod_{i=1}^{n} f(x_i \mid \theta) \tag{1}$$

where $f(x_i \mid \theta)$ is the probability density for the i^{th} independent response and k is a constant emphasizing the fact that the likelihood is a function of θ not a density for θ. The likelihood function is the key source of information to be drawn from a given model-data combination. Often the mode of the likelihood function $\hat{\theta}$, the maximum likelihood estimator, is the basis of frequentist inference. The local curvature of the log-likelihood about its mode provides the basis of the Fisher Information and related Cramer-Rao information lower bound.

The Bayesian approach or perspective is based on the joint posterior density which can be expressed as;

$$p(\theta \mid data) = c \cdot p(\theta) \cdot L(\theta \mid data) \tag{2}$$

here $p(\theta)$ is the prior density, $L(\theta \mid data)$ the likelihood function and c the constant of integration. All three functions of θ can be viewed as weighting the parameter space, with prior and posterior densities restricted to a probability scale.

The posterior density $p(\theta \mid data)$ can be viewed as an updated description of the researcher's beliefs regarding potential values of the parameter θ and is interpreted conditionally upon the observed data. From baseline beliefs for θ reflected in the shape of the prior density $p(\theta)$, the likelihood function updates these beliefs in light of the observed model and data giving the posterior density. Once the joint posterior is obtained, integration is employed in the Bayesian setting to obtain marginal posterior densities for any given θ_i. For example;

$$p(\theta_1 \mid data) = \int \cdots \int p(\theta \mid data) \mathrm{d}\theta_2 \mathrm{d}\theta_3 \cdots \mathrm{d}\theta_p \tag{3}$$

gives the marginal posterior for θ_1 alone. The central region of this density is a Bayesian credible region which can be used for estimation regarding θ_1. Both approaches to inference may employ approximation, typically based on larger sample sizes, to evaluate required tail areas or central estimation regions. With the advent of Markov Chain Monte Carlo (MCMC) based methods calculations in many Bayesian settings are possible [8].

Bayesian statistical analysis as an approach to the interpretation of statistical models has grown rapidly in application over the past several decades. This has been especially true in the basic sciences which, while traditionally not very open to the more subjective Bayesian perspective, have been open to its broader and more flexible modeling approach [13]. Bayesian analysis does however require an understanding of the analyst's set of prior beliefs regarding the set of population characteristics or parameters of interest and provides a process by which they will be updated by the observed model-data combination. Typically these are defined in the context of a mathematical model and beliefs must be assumed for the entire set of potential values

for the population parameters, even those that may not be significant in the final analysis.

The importance of the likelihood function is sometimes overlooked. It is the tool by which model and observed data are combined in both frequentist and Bayesian settings. As noted above, its properties underlie the Cramer-Rao information bound and in large samples it achieves a quadratic log-likelihood [1]. When viewed from a Bayesian perspective, the likelihood function updates initial preferences given by the prior density, giving new weighted preferences in the form of the posterior density.

Utility functions are also a concept that can be employed to model individual preferences regarding unknown parameter values when choices are to be made from a set of possible values. Their properties underlie much of economic consumption theory in regards to the individual consumer, production choices of the firm, and broader social utility issues [14]. In general, if available, utility functions can be used along with the posterior density to obtain an expected utility function that may be used to model consumer preferences.

Expected utility also has a long history in economic thought and provides a context for the study of preferences and related behavior [15]. It has also been viewed as a basis for Bayesian inference [16] [17]. In large samples it is possible to develop an expected utility interpretation in relation to the likelihood function itself, in relation to the processing and consumption of information. The effects of large samples in relation to expected utility have previously been examined from the perspective of laws of large numbers [18]. But in large samples the asymptotic shape of the log-likelihood function, if placed in a Bayesian setting, provides direct insight into the empirical support offered for specific values of population parameters. In large samples with non-informative priors, the log-concavity of the likelihood function yields the shape of the log-posterior density.

This can initially be seen in relation to central limit theorems. Subject to regularity conditions [1] the following result holds as $n \to \infty$,

$$\hat{\theta} \sim N\left(\theta, I^{-1}(\theta)\right) \tag{4}$$

a sampling theory result from the frequentist perspective where $I(\theta)$ is the Fisher information and $I^{-1}(\theta)$ the well known Cramer-Rao bound.

It is also true that, conditional on the data *x*, as $n \to \infty$;

$$\theta \mid x \sim N\left(\hat{\theta}, J^{-1}(\theta)\right) \tag{5}$$

from the Bayesian perspective with $J(\theta)$ the observed Fisher information. Note in large samples that $J^{-1}(\theta) \to I^{-1}(\theta)$ in probability.

The Bayesian perspective on statistics can be viewed as providing models for learning based behavior. The "prior" density $p(\theta)$ serves as an initial baseline for the analyst's beliefs regarding potential values of θ. The prior is then updated as observed data is processed. The information is collected via the likelihood function and processed through the prior-likelihood pair to give the posterior density. The result is a reweighting

of belief regarding θ.

3. Likelihood Related Stability in the Posterior Density

The learning aspect of Bayesian methods is based on the likelihood function. The information-theoretic aspects of the likelihood function summarize and provide information to update beliefs regarding θ. There are various approaches to assessing the rate and stability with which the posterior modifies or "learns". The inferential stability of the posterior density can be seen as a function of its rate of change and (on a logarithmic scale) depends directly on the additive rates of change in the prior density and log-likelihood.

Assuming a scalar θ we have;

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln p(\theta\,|\,\text{data}) &= \frac{\partial}{\partial\theta}\ln c + \frac{\partial}{\partial\theta}\ln p(\theta) + \frac{\partial}{\partial\theta}\ln L(\theta\,|\,\text{data}) \\ &= \frac{\partial}{\partial\theta}\ln p(\theta) + \frac{\partial}{\partial\theta}\ln L(\theta\,|\,\text{data})\end{aligned} \tag{6}$$

Note that Bayesian inference, by employing the likelihood function, inherits many optimal properties of the frequentist-likelihood approach to inference. This includes the score function, which is at the heart of frequentist-likelihood inference [3] and can be written;

$$S(\theta) = \frac{\partial}{\partial\theta}\ln L(\theta\,|\,\text{data}) \tag{7}$$

and is also a component of the posterior rate of change. The only difference between the rate of change of the log posterior and the score function is the rate of change in the log prior, which is zero if the prior is non-informative or constant.

$$\frac{\partial}{\partial\theta}\ln p(\theta\,|\,data) = \frac{\partial}{\partial\theta}\ln p(\theta) + S(\theta\,|\,\text{data}) \tag{8}$$

In these settings, the score function provides information regarding the percent rate of change in the posterior as a function of θ. In effect the elasticity of the Bayesian posterior density.

In large samples the derivative of the log-likelihood is also useful in describing the stability of probability preferences. Given a non-informative or constant prior and large sample it follows that;

$$\lim_{n\to\infty}\frac{\partial}{\partial\theta}\ln p(\theta\,|\,\text{data}) = \lim_{n\to\infty} S(\theta\,|\,\text{data}) \tag{9}$$

In other words the relative changes in the log-posterior will reflect directly the asymptotic behavior of the Score function.

Taking a frequentist perspective on the data, the asymptotic distribution of the score function can be applied to provide large sample bounds for the standardized rate of change in the log posterior in relation to the log prior baseline. Giving the result;

$$S(\theta\,|\,\text{data})\Big/\sqrt{J(\theta)} \sim N(0, I) \tag{10}$$

where I is the identity matrix here. In case of a scalar θ it follows that;

$$-2\sqrt{J(\theta)} \le \frac{\partial}{\partial\theta}\ln p(\theta|\text{data}) - \frac{\partial}{\partial\theta}\ln p(\theta) \le 2\sqrt{J(\theta)} \tag{11}$$

Thus on a logarithmic scale the difference in rates of change or elasticity in the posterior versus prior is bounded by the observed Fisher information $J(\theta;x)$. Thus the information provided by the likelihood function is key in assessing bounds on a measure of change from prior to posterior. Note that a similar result can be expressed more generally in terms of the Kullback-Liebler distance measure [19].

In multivariate parameter settings the effect of integrating out unwanted or nuisance parameters may affect the nominal accuracy of the resulting marginal posterior. Thus the similarity between the results that as $n\to\infty$, $\hat{\theta}_j \sim N\left(\theta_j, I^{-1}\left(\theta_j\right)\right)$ or $\theta_j \mid x \sim N\left(\hat{\theta}_j, J^{-1}\left(\theta_j\right)\right)$ may not be as direct on a marginal scale.

4. Utility Functions

Utility theory has a long history and can be found in its most developed form in economic theory. Utility functions themselves define a preference relation. The work of Von Neumann [20] and Samuelson [21] and Arrow [22] provided axioms for the definition and application of utility functions in relation to expected utility. If the axioms are satisfied, the individual is said to be rational and preferences can be represented by a utility function.

The Von Neumann-Morgenstern utility representation theorem [14] has four possible axioms, though the independence axiom is sometimes dropped and is so here where a simple scalar parameter setting is examined. Where the large sample log-likelihood $l(\theta) = l(\theta \mid \text{data})$ is concave (quadratic) and continuous, the axioms can be seen to apply directly with the ranking $A \ge B$ defined by $l(A) \ge l(B)$, and A, B and C values for θ in the support of the log-likelihood function;

1) Completeness: The individual either prefers A to B, or is indifferent between A and B, or prefers B to A. The concave and continuous weighting provided by the large sample shape of the log-likelihood function satisfies this condition.

2) Transitivity: For every A, B and C with $A \ge B$ and $B \ge C$ we must have $A \ge C$. This follows directly from the continuous, quadratic concave shape of the large sample log-likelihood function.

3) Continuity: Let A, B and C be such that $A \ge B \ge C$; then there exists a probability weighting p such that B is equally good as $pA + (1-p)C$. This holds for the continuous, quadratic concave shape of the large sample log-likelihood when weighted by an appropriately chosen prior density $p = p(\theta)$.

This formality is useful when incorporating probability in relation to utility or expected utility. That said, it is often the case in large samples that central limit theorems, the weak law of large numbers and strong law of large numbers apply and modify the determination of probability based preferences and related values of expected utility. This is discussed in [18] from a non-likelihood based large sample frequentist perspective.

In large samples the log-likelihood (or likelihood) provides a pseudo-utility function

that satisfies the above axioms of utility in relation to preferred values for the parameter θ. The likelihood and prior density can be interpreted as a conceptual pair and the resulting posterior or log-posterior an expected utility providing a preference related weighting of the parameter space.

The log-likelihood converges to a quadratic form across a wide set of assumed probability models for the observed data. These typically comprise the exponential family of probability densities. The associated regularity conditions can be found in [1].

The log-concavity of the large sample likelihood can be expressed;

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta \mid y) < 0 \tag{12}$$

While initially most likelihood functions may not have the properties necessary to be viewed as utility functions, in large samples many likelihoods do have these properties, subject to regularity conditions, and are log-concave, continuous and differentiable.

In the scalar θ case the result;

$$\hat{\theta} \sim N\left(\theta, I^{-1}(\theta)\right) \tag{13}$$

can be interpreted as the likelihood function having a large sample bell curve shape. The related log-likelihood function is quadratic, an acceptable form for consideration as a utility function and the required conditions above are met.

5. Interpreting Risk Aversion in Large Samples

As the large sample pairing of prior probability and likelihood allows for an expected utility perspective on the resulting posterior density function, work by Jeffrey [23] can be applied. This emphasizes taking probability and utility functions as pairs in relation to developing optimal probability based preferences. Here the (prior, likelihood) pair, processed through Bayes Theorem, provides a large sample expected utility function; the posterior or log-posterior distribution, for ranking preferences regarding θ.

If the collection and interpretation of data in relation to an assumed parametric model is viewed as a process of consuming information, and incorporating probability based preferences based on likelihood functions to update existing belief, then measures of risk aversion related to expected utility can be applied in a general information context.

The Arrow-Pratt absolute risk aversion (ARA) measure [14] is defined generally as;

$$ARA(w) = -\frac{u''(w)}{u'(w)} \tag{14}$$

where $u(w)$ is an expected utility function. It is often used as a standardized measure of risk aversion in regard to expected utility. In a large sample where we take a relatively flat prior density, the ARA measure is simply the inverse of the standardized score function and thus a standardized rate of preference modification with regard to the posterior density function and the parameter θ.

Writing $u(\theta) = \ln p(\theta \mid data)$ as the log posterior density and assuming a first order

condition on the prior density $\frac{\partial}{\partial\theta}\ln p(\theta)=0$, we can express the ARA measure in terms of;

$$\begin{aligned} ARA(\theta)^{-1} &= \frac{u'(\theta)}{-u''(\theta)} = \frac{\frac{\partial}{\partial\theta}\ln p(\theta\,|\,\text{data})}{-\frac{\partial^2}{\partial\theta^2}\ln p(\theta\,|\,\text{data})} \\ &= \frac{\frac{\partial}{\partial\theta}\ln c+\frac{\partial}{\partial\theta}\ln p(\theta)+l'(\theta\,|\,\text{data})}{-\frac{\partial^2}{\partial\theta^2}\ln c-\frac{\partial^2}{\partial\theta^2}\ln p(\theta)-l''(\theta\,|\,\text{data})} \\ &= \frac{l'(\theta\,|\,\text{data})}{-l''(\theta\,|\,\text{data})} = \frac{S(\theta\,|\,\text{data})}{J(\theta)}. \end{aligned} \tag{15}$$

This interpretation also allows for a central limit theorem related argument regarding ARA^{-1} in large samples which bounds the ARA measure of risk aversion in relation to the Cramer Rao information bound;

Theorem 1 *The function* $ARA(\theta)^{-1}=\frac{S(\theta\,|\,\text{data})}{J(\theta)}$ *has a large sample* $N\left(0,I(\theta)^{-1}\right)$ *distribution for large n.*

Proof. Taking limits and assuming standard likelihood related regularity conditions hold, the central limit theorem for the score function, the strong law of large numbers and Slutsky's theorem can be applied giving;

$$ARA^{-1}=\frac{S(\theta\,|\,\text{data})}{\sqrt{J(\theta)}}\cdot\frac{1}{\sqrt{J(\theta)}}\to N(0,1)\cdot\frac{1}{\sqrt{I(\theta)}}\sim N\left(0,I(\theta)^{-1}\right). \tag{16}$$

This is a simple restatement of the large sample or asymptotic efficiency of the score function and optimality of the Cramer-Rao lower bound [1], but in relation to the consumption of information and related risk aversion. This provides an asymptotic variance for ARA^{-1} when appropriate. It can also be argued that the ARA^{-1} measure is efficient in the processing of information as its variation attains the Cramer-Rao information bound in large samples.

While not practical, a large sample 95% confidence related bound on ARA^{-1} or ARA can be defined in relation to statistical information;

$$\begin{aligned} -2\sqrt{1/I(\theta)} \le ARA^{-1} \le 2\sqrt{1/I(\theta)} \\ (1/2)\sqrt{I(\theta)} \le ARA \le -(1/2)\sqrt{I(\theta)} \end{aligned} \tag{17}$$

It is interesting to note that the Likelihood Principle is implicitly relevant to this result. As noted earlier, this principle states that inference from two proportional likelihood functions, $L_1(\theta\,|\,x)=c\cdot L_2(\theta\,|\,x)$, should be the same. This is equivalent to saying that the derivative of the log-likelihoods or Score functions are equivalent $S_1(\theta\,|\,\text{data})=S_2(\theta\,|\,\text{data})$ and thus the rate of learning is equivalent if the priors in question are non-informative.

In terms of utility, this implies that two proportional log-likelihood functions in large

samples can be viewed as having identical large sample ARA^{-1} values in relation to the information content of the respective model-data combinations. Thus proportional likelihoods yield similar levels of risk aversion in large samples.

Note that the likelihood function is log-concave generally when we have the condition;

$$\log L\left(t\theta_1+(1-t)\theta_2 \mid y\right) > t\log L\left(\theta_1 \mid y\right)+(1-t)\log L\left(\theta_2 \mid y\right). \tag{18}$$

This may hold in some small sample settings with non-informative prior densities. The simplest approach to ensuring a log-concave likelihood in small samples is to work with log-concave densities [24]. This reflects the basic property that If X and Y have log-concave densities, so does $X+Y$. The Normal, Poisson and Binomial distributions for example have this property, Note that the Cauchy, Pareto and log-Normal distributions are not log concave densities. Mixtures of the normal and other distributions may or may not have this property.

6. Prior Selection: Enabling Likelihood Based Learning

As noted above, Bayesian methods obtain their accuracy and informative nature by depending heavily on the likelihood function. In emphasizing a learning model perspective, and imposing the requirement that we learn from the likelihood function, the technical link between posterior and likelihood allows for consideration of the likelihood in relation to choosing a prior. In particular, this can be examined from the perspective of statistical information and linking aspects of the log-likelihood with posterior stability, matching the curvature of the log-likelihood function, the observed Fisher Information, to the curvature of the posterior density. This gives rise to conditions that help guide the selection of prior densities.

The Bayesian perspective reflects a learning process in regard to the parameter θ. This learning process should not be a function of pre-existing belief which in a sense sets the baseline of existing knowledge. Rather it should reflect the properties and information of the model-data combination in the form of the likelihood function. Here we suggest an approach to prior selection which focuses on matching the information properties of the likelihood and posterior densities and gives a family of prior densities from which to choose.

Define the concept of *posterior information* as the local curvature of the log-posterior about its mode;

$$\begin{aligned} -\frac{\partial^2 \ln p(\theta \mid \text{data})}{\partial\theta^2} &= -\frac{\partial^2}{\partial\theta^2}\left[\ln c+\ln p(\theta)+\ln L(\theta \mid \text{data})\right] \\ &= -\frac{\partial^2}{\partial\theta^2}\ln p(\theta)-\frac{\partial^2}{\partial\theta^2}\ln L(\theta \mid \text{data}) \\ &= -\frac{\partial^2}{\partial\theta^2}\ln p(\theta)+J(\theta) \end{aligned} \tag{19}$$

where $J(\theta)$ is the observed Fisher information $J(\theta)=-\frac{\partial^2}{\partial\theta^2}\ln L(\theta \mid \text{data})$.

Given the selection of a prior which is to be non-informative at the level of information processing, and assuming that standard regularity conditions apply to the likelihood function [1], we set the following second order condition on the prior density;

$$\frac{\partial^2}{\partial\theta^2}\ln p(\theta)=0 \tag{20}$$

or more reasonably;

$$\frac{\partial^2}{\partial\theta^2}\ln p(\theta)=k \tag{21}$$

where k is a constant. This implies that, up to a multiplicative constant, the likelihood based Fisher Information in the model-data combination is the basis of all Bayes posterior information. Researchers learn from the likelihood, not from the prior.

The family of *information similar* priors chosen in this manner are non-informative to the second order and are of the form;

$$\ln p(\theta)=a\theta^2+b\theta+d \tag{22}$$

where a,b,d are constants. Note that this implies an exponential family related class of prior distributions from which to choose, which may or may not be conjugate to the posterior density. A reasonable restriction on the constants a,b,d is to require the prior to be well defined. The normal, binomial and Poisson distributions satisfy this restriction as do most standard choices for priors. Technically so do flat or highly non-informative priors. Distributions with third or higher order polynomials are ruled out.

Some examples of priors that are not acceptable in this setting include;

$$\begin{aligned} &p(\theta)\propto\exp\left(\theta^3\right)\\ &p(\theta)\propto\exp\left(\theta^m\right),m>3\\ &p(\theta)\propto\exp\left(\sin\left(g(\theta)\right)\right) \end{aligned} \tag{23}$$

Note that while focusing here on learning from likelihood, the effect of integration or shrinkage may imply some prior effect in the multivariate setting when integrating to obtain marginal posteriors. The use of hyperparameters in hierarchical or empirical Bayesian settings raise related issues. These are examined in detail elsewhere.

The Jeffreys prior [25] in large samples achieves such an information similar effect. Considering the Bayesian asymptotic result $\theta\sim N\left(\hat{\theta},J^{-1}(\theta)\right)$, the Jeffreys prior can be taken as the inverse of the observed variance or Cramer-Rao bound $\left(J^{-1}(\theta)\right)^{-1}$. This is essentially the inverse of the local curvature of the log likelihood function and will behave locally as the inverse of asymptotic variation; it will be relatively flat where the likelihood is pronounced. This will be approximately non-informative in the sense defined here; it focuses on preserving the local shape of the likelihood about its mode as a key element of the shape of the posterior density.

Example

Consider the case of nonlinear regression with Normal error.

$$y_i = \eta(x_i;\beta) + \varepsilon_i \tag{24}$$

where the x are fixed, the ε are $i.i.d.\ N(0,\sigma^2)$ and $\eta(\cdot)$ represents the nonlinear regression surface. Let β be a scalar parameter. The Jeffreys prior for this was suggested in [26] and is given by;

$$p(\beta) = \frac{F(x;\beta)' F(x;\beta)}{\sigma^2} \tag{25}$$

where $F(\cdot)$ is the first order derivative of $\eta(x_i;\beta)$ with regard to β, where σ is assumed known or estimated by the *MSE*. This is acceptable in terms of information similarity if it has the property;

$$\frac{\partial^2}{\partial\beta^2}\left[\ln\left(\frac{F(x;\beta)' F(x;\beta)}{\sigma^2}\right)\right] = k \tag{26}$$

This implies that nonlinear regression surfaces should not be too complex as a function of β if they enter into the related processing of likelihood based information.

Multiparameter Settings

In multiparameter settings, where $\theta = (\theta_1,\cdots,\theta_p)$, this approach provides guidance in selecting priors if information matching is applied. The resulting conditions are given by;

$$\begin{aligned}
&\frac{\partial^2}{\partial\theta_1^2}\ln p(\theta) = k_1\\
&\frac{\partial^2}{\partial\theta_2^2}\ln p(\theta) = k_2\\
&\qquad\vdots\\
&\frac{\partial^2}{\partial\theta_p^2}\ln p(\theta) = k_p.
\end{aligned} \tag{27}$$

where k_i are constants. If symmetry or independence is useful, $p(\theta) = \prod_{j=1}^{p} p_j(\theta_i)$ can be assumed. With cross derivatives set equal to zero, multiparameter priors can be taken with the general form;

$$\ln p(\theta) = \sum a_i\theta_i^2 + \sum b_i\theta_i + \sum d_i \tag{28}$$

This rules out multivariate prior distributions with factors of θ that are higher order polynomials of θ or transcendental functions such as

$$\ln p(\theta) \propto \left(\sin(\theta_1),\cdots,\sin(\theta_p)\right) \tag{29}$$

or

$$p(\theta) \propto \left(\exp(\exp(\theta_1)),\cdots,\exp(\exp(\theta_p))\right). \tag{30}$$

This approach to prior selection can be seen as imposing the log-concavity of the

likelihood function, which yields a log-concave joint posterior density and risk averse behavior as the amount of information increases. Note that the approach given by reference priors [2], also reflect the idea of selecting priors to maximize the amount learned, but typically averaged over the sample space. A formal Bayesian conditional perspective reflecting the observed data is maintained here.

7. Discussion

This paper reviews and develops links between several concepts; large sample likelihood, expected utility, risk aversion, posterior stability and aspects of prior selection. These are broadly defined concepts providing templates for the organization and study of behavior and how such behavior is modified in the light of information. In relation to the utility and expected utility aspect, it is information itself that is the consumed good of interest. In the context of a particular large sample model-data combination the Fisher Information and Cramer-Rao information bound are directly related to measures of expected utility based risk aversion.

In large samples the concavity of the log-likelihood of the asymptotic normal density allows for use of the likelihood function in relation to the concept of utility and expected utility. The Cramer Rao information bound is seen to have a large sample relationship in providing bounds on the elasticity of the posterior density and the Arrow-Pratt measure of risk aversion. The imposition of information similarity on the likelihood-posterior relationship provides direct application of the Fisher Information from a learning model perspective. It provides a class of information similar prior densities that emphasize likelihood as the source of model-data related information.

To summarize, the likelihood function is a key element in the processing of information through defined model-data constructs. This is true from various perspectives. The possible use of the large sample likelihood function as a utility function itself allows for the linking of concepts of risk aversion, as expressed by the Arrow-Pratt measure, with statistical information. As well, the implicit learning oriented focus of the Bayesian perspective, if focused on the properties of the large sample likelihood, leads to restrictions on the type of priors available when information from both Bayesian and frequentist perspectives directly reflect the Fisher information.

References

[1] Casella, G. and Berger, R.L. (2002) Statistical Inference. 2nd Edition, Duxbury Press, Pacific Grove.

[2] Bernardo, J.M. and Smith, A.F.M. (1994) Bayesian Theory. John Wiley and Sons Inc., New York. http://dx.doi.org/10.1002/9780470316870

[3] Sims, C.A. (2000) Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples. *Journal of Econometrics*, **95**, 443-462. http://dx.doi.org/10.1016/S0304-4076(99)00046-9

[4] Pawitan, Y. (2001) In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford Science Publications, Clarendon Press, Oxford.

[5] Pierce, D.A. and Peters, D. (1994) Higher-Order Asymptotics and the Likelihood Principle:

One Parameter Models. *Biometrika*, **81**, 1-10. http://dx.doi.org/10.1093/biomet/81.1.1

[6] Frieden, B.R. (2004) Science from Fisher Information: A Unification. Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/CBO9780511616907

[7] Akaike, H. (1981) Likelihood of a Model and Information Criteria. *Journal of Econometrics*, **16**, 3-14. http://dx.doi.org/10.1016/0304-4076(81)90071-3

[8] Gilks, W.R., Richarson, S. and Spiegelhalter, D.J. (1996) Markov Chain Monte Carlo in Practice. Chapman and Hall, New York.

[9] Aeschbacher, S., Beaumont, M.A. and Futschik, A. (2012) A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation. *Genetics*, **192**, 1027-1047. http://dx.doi.org/10.1534/genetics.112.143164

[10] Luo, R., Hipp, A.L. and Larget, B. (2007) A Bayesian Model of AFLP Marker Evolution and Phylogenetic Inference. *Statistical Applications in Genetics and Molecular Biology*, **88**, 1813-1823.

[11] Berger, J.O. (1990) Robust Bayesian Analysis: Sensitivity to the Prior. *Journal of Statistical Planning and Inference*, **25**, 303-328. http://dx.doi.org/10.1016/0378-3758(90)90079-A

[12] Datta, G.S., Mukerjee, R., Ghosh, M. and Sweeting, T.J. (2000) Bayesian Prediction with Approximate Frequentist Validity. *Annals of Statistics*, **28**, 1414-1426.

[13] Lau, M.S.Y., Marion, G., Streftaris, G. and Gibson, G. (2015) A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLOS Computational Biology*, **11**, e1004633. http://dx.doi.org/10.1371/journal.pcbi.1004633

[14] Varian, H.R. (1992) Microeconomic Analysis. 3rd Edition, W.W. Norton & Company, New York.

[15] Anand, P. (1993) Foundations of Rational Choice under Risk. Oxford University Press, Oxford.

[16] Savage, L. (1962) Foundations of Statistical Inference: A Discussion. Methuen, London.

[17] Good, I.J. (1984) A Bayesian Approach in the Philosophy of Inference. *British Journal for the Philosophy of Science*, **35**, 161-166. http://dx.doi.org/10.1093/bjps/35.2.161

[18] Feller, W. (1968) An Introduction to Probability Theory and Its Applications. Vol. 1, 3rd Edition, John Wiley and Sons, New York.

[19] Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**, 79-86. http://dx.doi.org/10.1214/aoms/1177729694

[20] Von Neumann, J. and Morgenstern, O. (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton.

[21] Samuelson, P. (1948) Consumption Theory in Terms of Revealed Preference. *Econometrica*, **15**, 243-253. http://dx.doi.org/10.2307/2549561

[22] Arrow, K.J. (1971) The Theory of Risk Aversion. In: Helsinki, Y.J.S., Ed., *Aspects of the Theory of Risk Bearing*, Reprinted in Essays in the Theory of Risk Bearing, Markham Publ. Co., Chicago, 90-109.

[23] Jeffrey, R. (1983) The Logic of Decision. 2nd Edition, University of Chicago Press, Chicago.

[24] Dumbgen, L. and Rufibach, K. (2009) Maximum Likelihood Estimation of a Log-Concave Density and Its Distribution Function: Basic Properties and Uniform Consistency. *Bernoulli*, **15**, 40-68. http://dx.doi.org/10.3150/08-BEJ141

[25] Jeffreys, H. (1961) Theory of Probability. Oxford University Press, Oxford.

[26] Eaves, D.M. (1983) On Bayesian Nonlinear Regression with an Enzyme Example. *Biometrika*, **70**, 373-379. http://dx.doi.org/10.1093/biomet/70.2.373