NORTHWESTERN UNIVERSITY

A Bayesian Approximation to the Integrated Likelihood Function with Applications in Meta-Analysis

A DISSERTATION PROSPECTUS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Timothy Ruel

EVANSTON, ILLINOIS

September 2023

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Table of Contents

CHAPTER 1

Preface

1.1. Introduction

The research for my dissertation involves developing a novel algorithm for numerically integrating the likelihood function of a statistical model with respect to a nuisance parameter. This prospectus aims to demonstrate how the algorithm works and explain the appeal of using an integrated likelihood function over other types of pseudolikelihood functions to make inferences about the parameter of interest in a model.

1.2. Motivation

The motivation behind my research developed out of the observation that the expression for the integral of a likelihood function follows a form similar to that of the calculation of the marginalizing constant of a posterior distribution.

CHAPTER 2

Background

2.1. Assumptions

Consider a random sample $\mathbf{x} = (x_1, ..., x_n)$ drawn from a population. What can we say about the population based on $\mathbf{x}$? Where is its point of central tendency located? Are its values clustered tightly around this point, or are they more diffuse? Are they distributed symmetrically or skewed to one side or the other? Questions like these were the original motivation behind the field of statistical inference, and many of the techniques devised to answer them are still used by statisticians today.

It is important to remember, however, that the real world is messy and no mathematical function will ever perfectly capture the complexities of a population or random process whose properties we wish to infer. To overcome this difficulty, statisticians sacrifice a small amount of accuracy for (hopefully) a large reduction in complexity by imposing additional assumptions on the population of interest. These assumptions are essentially never true in the sense that they are not a flawless representation of reality, but they may nevertheless serve as convenient approximations capable of producing sufficiently accurate answers in their own right. As George Box famously put it, "All models are wrong, but some are useful."

2.1.1. Data-Generating Mechanism

And indeed, in the aggregate these assumptions create what is known as a statistical model. In its most general framework, a statistical model can be formulated as a tuple $(\mathcal{S}, \mathcal{P})$ where $\mathcal{S}$ is the set of all possible observations (i.e. the population), and $\mathcal{P}$ is a set of probability distributions on $\mathcal{S}$. The first and (in this author's opinion) the most fundamental assumption we make when defining our models is that there exists some unknown mechanism in the population that generates the data we observe from $\mathcal{S}$. This mechanism is what induces the "true" probability distribution on $\mathcal{S}$ though $\mathcal{P}$ need not contain this distribution, and in practice it seldom does.

2.1.2. Parameter Existence

Another assumption found in almost every model is that the set $\mathcal{P}$ is considered to be *parameterized*. That is, we assume the probability distributions contained in $\mathcal{P}$ are indexed by a *parameter* that controls their features.[1] This parameter acts like a tuning dial for the population - rotate the dial and certain behaviors of the population (e.g. its location, scale, or shape) will change. Much of statistical inference can be boiled down to figuring out the particular value to which a population's dial has been set. We will denote this assumption as $\mathcal{P} = \{\mathcal{P}_\theta | \, \theta \in \Theta\}$, where θ denotes the parameter, and Θ, the parameter space, represents the set of all possible values θ can take on.[2]

Statisticians also like to assume the parameters in their models can be uniquely identified based on the data they observe. A model is considered *identifiable* if having perfect knowledge of the population enables us to determine the true value of its parameter with absolute certainty.[3] More formally, for any two parameters θ_1 and θ_2 in Θ, if $\mathcal{P}_{\theta_1} = \mathcal{P}_{\theta_2}$, then it must follow that $\theta_1 = \theta_2$. A model that is not identifiable could potentially have two or more distinct parameter values that give rise to the same probability distribution. Since we have already assumed $\mathcal{P}$ is the mechanism generating the data we have observed in the first place, this would make it impossible to determine which value is the "correct" one on the basis of the data alone. Statisticians impose the identifiability criterion on their models as a means of avoiding this undesirable situation.

2.1.3. Parameter Space Dimension

The dimension of the parameter space Θ is another critical decision statisticians must make when choosing the best model for their research. The most frequent choice is for Θ to be of finite dimension. That is, $\Theta \subseteq \mathbb{R}^k$, where $k \in \mathbb{Z}^+$. Models that satisfy this assumption are said to be *parametric*. Common examples of parametric models include the normal family of distributions as well as the Poisson family.

---

[1] We will consider the phrases "parameter", "population parameter", and "model parameter" all to have the same meaning in this paper and use them interchangeably.

[2] Note that θ can be and in fact usually is a multi-dimensional vector whose components represent various sub-parameters of the population.

[3] This is of course almost always impossible in practice, but in theory it could be accomplished by obtaining an infinite number of observations from $\mathcal{S}$ or simply all of its observations if $|\mathcal{S}|$ is finite.

In contrast, a *nonparametric model* is one in which $\Theta \subseteq V$, where V is an infinite-dimensional space. The name is a bit of a misnomer in the sense that nonparametric models do not actually lack parameters, but rather they are flexible regarding the exact number and properties of the parameters they do have.

Finally, *semiparametric* models are those whose parameter spaces have components of both finite and infinite dimensionality. That is, $\Theta \subseteq \mathbb{R}^k \times V$, where again V is an infinite-dimensional space. Usually it is only the finite-dimensional component of the parameter in which we are interested while the infinite-dimensional component is considered a nuisance parameter.

For the purposes of this paper, we will consider only models that consist of a pair $(\mathcal{S}, \mathcal{P})$ such that the probability distributions in $\mathcal{P}$ are parameterized and identifiable, and there exists a "true" probability distribution inducing the data-generating process from $\mathcal{S}$, though this distribution is not necessarily contained in $\mathcal{P}$. In addition, while estimation of the parameters in nonparametric and semiparametric models is also a major topic of interest in statistical inference, we will restrict our attention here solely to parametric models.

2.2. The Likelihood Function

Once we have chosen a model $(\mathcal{S}, \mathcal{P})$, our goal becomes to identify the "true" distribution in $\mathcal{P}$ or, failing that, the one that best approximates the truth. Since we have assumed our model is parametric and identifiable, this is equivalent to making inferences about the value of the k-dimensional parameter θ indexing the distributions in $\mathcal{P}$ on the basis of some data we observe. Classically, these inferences come in the form of point estimates, interval estimates, or hypothesis tests though other techniques exist as well. A sensible choice to use as an estimate for the value of θ is one which causes the data actually observed to have the highest possible *post-hoc* probability of occurrence out of all possible values in Θ. To formalize this notion, we need some way of analyzing the joint probability of our sample as a function of our parameter θ.

2.2.1. The Discrete Case

Suppose X is a discrete random variable with probability mass function $p(x;\theta)$. For a single observation $X = x$, the *likelihood function* for θ is defined as

$$L(\theta) \equiv L(\theta; x) = p(x;\theta), \ \ \theta \in \Theta. \tag{2.2.1}$$

That is, when our sample consists only of a single observation, the likelihood function for θ is simply equal to $p(x;\theta)$ itself. However, while $p(x;\theta)$ is viewed as a function of x for fixed θ, the reverse is actually true for $L(\theta;x)$; we view it as a function of θ for fixed x. The positioning of the arguments θ and x is a reflection of this difference in perspectives.

In this case, we may interpret $L(\theta)$ as the probability that $X = x$ given that θ is the true parameter value. It is crucial to note that this is *not* equivalent to the inverse probability that θ is the true parameter value given $X = x$. Though intuitively appealing, this interpretation constitutes a fundamental misunderstanding of what a likelihood function is, and great care must be taken to avoid it.

This definition be extended to include the more common scenario in which the sample consists of multiple observations. For a sample of size n taken from our sample space $\mathcal{S}$, the likelihood is defined as

$$L(\theta; x_1, ..., x_n) = p(x_1, ..., x_n; \theta), \ \ \theta \in \Theta. \tag{2.2.2}$$

That is, it is equal to the joint probability of the observations $x_1, ..., x_n$, considered as a function of θ. When the observations are independent and identically distributed, we can further express the likelihood as

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i;\theta), \ \ \theta \in \Theta. \tag{2.2.3}$$

2.2.2. The Continuous Case

When X is instead a continuous random variable, the likelihood for θ may still be defined as it is in Equation 2.2.1 and Equation 2.2.2. Note that $p(x;\theta)$ has switched from being a probability *mass* function to a probability *density* function over the support of X. We must therefore forfeit our previous

direct interpretation of $L(\theta)$ as a probability since $p(x;\theta)$ no longer represents $\mathbb{P}(X = x|\theta)$. We may still think of the likelihood as being proportional to the probability that X is "close" to the value x though.[4] Specifically, for two different samples x_1 and x_2, if $L(\theta; x_1) = c \cdot L(\theta; x_2)$, where $c > 1$, then under this model we may conclude X is c times more likely to assume a value closer to x_1 than x_2 given that θ is the true value of the parameter.

As in the discrete case, we must also be careful here to avoid using $L(\theta)$ to make direct statements of probability about θ. Indeed, despite our use of one in its definition, the likelihood function is *not* itself a probability density over the parameter θ and need not obey the same laws as one.

2.2.3. Maximum Likelihood Estimation

Maximum likelihood estimation is one of the most powerful and widespread techniques for obtaining point estimates of model parameters based on some observed data x. The original intuition behind the method derives from the observation that when faced with a choice between two possible values of a parameter, say θ_1 and θ_2, the sensible choice is the one that makes the data we did observe more probable to have been observed. Fortunately, we have already defined the likelihood function as a means of capturing this probability, which makes expressing this decision rule in terms of it very easy - we simply choose for our estimate the option that produces the higher value of the likelihood function. That is, if $L(\theta_1; x) > L(\theta_2; x)$, then θ_1 is the better estimate of the true parameter value and vice versa.

This can be extended to include as many parameter values as we would like. For n potential estimates of the true parameter, the best is the one that corresponds to the highest value of the likelihood function based on the observed data x. Taking this logic to its natural conclusion, the *maximum likelihood estimate* (MLE) of the parameter θ, which we will denote by $\hat{\theta}$ (pronounced "theta hat"), is the one that maximizes the value of the likelihood function among all possible choices of θ in the parameter space Θ. Formally,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta; x). \tag{2.2.4}$$

[4]Here, "close" means that that X is within a tiny neighborhood of x.

There is no singular method for finding the maximum likelihood estimate of a parameter. However, when the likelihood function is differentiable, it is often possible to calculate the MLE analytically using the derivative test for locating the local maxima of a function. In such cases, the MLE can be found by finding the value of θ that makes the derivative of the likelihood function with respect to θ vanish. A popular technique when finding this value is to take the natural logarithm of the likelihood first. This transformation is common enough that it has its own name - the *log-likelihood function*. Formally, it is defined as

$$\ell(\theta) \equiv \ell(\theta; x) = \log L(\theta; x), \;\; \theta \in \Theta. \tag{2.2.5}$$

When working with $\ell(\theta)$ instead of $L(\theta)$, any products in the latter have been transformed into sums in the former, making derivative calculations more tractable while still preserving the argument that corresponds to the global maximum, if it exists, of $L(\theta)$.

2.2.4. Model Parameter Decomposition

It is often the case that we are not interested in estimating the full parameter $\theta \in \Theta \subseteq \mathbb{R}^k$, but rather a sub-parameter ψ taking values in a set $\Psi \subseteq \mathbb{R}^m$, where $m < k$. In such an event, we refer to ψ as the *parameter of interest*.

2.2.5. Explicit Parameters

Consider first the case in which ψ is a sub-vector of θ, so that all the components of ψ are also components of θ. Then there exists a set $I = \{I_1, ..., I_m\} \subsetneq \{1, ..., k\}$ such that $\psi = (\psi_1, ..., \psi_m) = (\theta_{I_1}, ..., \theta_{I_m})$. We may further group the components of θ that are not a part of ψ into their own sub-vector, which we will refer to as the *nuisance parameter* and denote by $\lambda \in \Lambda \subseteq \mathbb{R}^{k-m}$. Specifically, let $N = \{N_1, ..., N_{k-m}\} \subsetneq \{1, ..., k\}$ such that $I \cup N = \{1, ..., k\}$ and $I \cap N = \emptyset$. Then $\lambda = (\lambda_1, ..., \lambda_{m-k}) = (\theta_{N_1}, ..., \theta_{N_{k-m}})$. θ can therefore be decomposed as $\theta = (\psi, \lambda)$, provided we shuffle the indices appropriately. In this case, ψ and λ are referred to as an *explicit* parameter of interest and nuisance parameter, respectively.

Nuisance parameters are so named for their ability to complicate inference regarding the parameter of interest. Despite not being the object of study themselves, they nevertheless are capable of modifying the distributions of our observations and therefore must be accounted for. The process by which this is accomplished is often nontrivial and indeed can constitute a significant barrier that must be overcome. For example, suppose we are interested in estimating the mean of a random variable Y, where $Y \sim N(\mu, \sigma^2)$. The full model parameter is $\theta = (\mu, \sigma^2)$ but since we are only interested in estimating the mean, the parameter of interest is $\psi = \mu$ and the nuisance parameter is $\lambda = \sigma^2$.

2.2.6. Implicit Parameters

Now consider the case in which the parameter of interest is the output of a function $\varphi : \Theta \to \Psi$. That is, $\psi = \varphi(\theta)$.[5] Note that Ψ is still assumed to be a subset of $\mathbb{R}^m$ where m is less than k, the dimension of the full parameter space Θ. This reduction in dimension implies the existence of a nuisance parameter $\lambda \in \Lambda$, where $\dim(\Lambda) = k - m$. However, there is no guarantee that a closed form expression exists for λ in terms of the original components of theta. We will refer to ψ and λ as being *implicit* parameters in this case.

2.3. The Pseudolikelihood Function

The construction of what is known as a *pseudolikelihood function* has long been the statistician's method of choice for eliminating nuisance parameters from a model. Suppose our parameter of interest is $\psi = \varphi(\theta)$ for some function ϕ and full model parameter $\theta \in \Theta$. If we let $\Theta(\psi) = \{\theta \in \Theta \,|\, \varphi(\theta) = \psi\}$, then corresponding to $\psi \in \Psi$ is the set of likelihoods $\mathcal{L}_\psi = \{L(\theta) \,|\, \theta \in \Theta(\psi)\}$. Simply put, a pseudolikelihood function is a summary of the values in $\mathcal{L}_\psi$ that does not depend on λ. There exist a variety of methods to obtain this summary but among the most popular are maximization, conditioning, and integration, each with respect to the nuisance parameter. We will explore each of these methods in more detail in the sections to come.

---

[5]Note that the first case is really just a special case of this second one in which $\varphi(\theta) = \varphi(\theta_1, ..., \theta_k) = (\theta_{I_1}, ..., \theta_{I_m})$, where $\{I_1, ..., I_m\}$ is the subset of the indices of the components of θ that also belong to ψ.

2.3.1. The Profile Likelihood

The profile likelihood is the most straightforward method for eliminating a nuisance parameter.

2.3.2. The Conditional Likelihood

2.3.3. The Marginal Likelihood

2.3.4. The Integrated Likelihood

2.4. The Appeal of the Integrated Likelihood Function

The appeal of the integrated likelihood function as a means of eliminating nuisance parameters from the model is that it incorporates

CHAPTER 3

Approximating the Integrated Likelihood

3.1. The Zero-Score Expectation Parameter

Let $\psi = \varphi(\theta)$ and λ denote the parameter of interest and nuisance parameter, respectively, for some statistical model $(\mathcal{S}, \mathcal{P}_\theta)$. Then the general expression to obtain an integrated likelihood for ψ may be written as

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda, \tag{3.1.1}$$

where $\pi(\lambda|\psi)$ is a conditional prior density for λ given ψ.

Severini (2007) considered the problem of selecting $\pi(\lambda|\psi)$ such that when the likelihood function is integrated with respect to this density, the result is useful for non-Bayesian inference. To do this, he outlined four properties (see Appendix A) that an integrated likelihood function must satisfy if it is to be of any use. He went on to prove that an integrated likelihood satisfying these properties could be obtained by first constructing a new nuisance parameter $\phi \in \Phi$ that is unrelated to the parameter of interest (in the sense that its maximum likelihood estimator remains roughly unchanged for all values of ψ) and then choosing a prior density $\pi(\phi)$ that is independent of ψ. Once chosen, the desired integrated likelihood function for ψ is given by

$$\bar{L}(\psi) = \int_\Phi \tilde{L}(\psi, \phi)\pi(\phi)d\phi, \tag{3.1.2}$$

where $\tilde{L}(\psi, \phi)$ is the likelihood function for the model after it has been reparameterized in terms of ϕ. It is important to note that the exact choice of prior density for ϕ is not particularly important; the only restriction we place upon it is that it must not depend on ψ.

Suppose that we have an explicit parameter of interest and nuisance parameter, so that $\theta = (\psi, \lambda)$. Then Severini (2007) defines this new nuisance parameter ϕ as the solution to the equation

$$\mathbb{E}(\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0)\Big|_{(\psi_0,\lambda_0)=(\hat{\psi},\phi)} = 0, \tag{3.1.3}$$

where $\ell_\lambda(\psi, \lambda) = \frac{\partial \ell(\psi,\lambda)}{\partial \lambda}$, ψ_0 and λ_0 denote the true values of ψ and λ, and $\hat{\psi}$ is the MLE for ψ_0. In other words, for a particular value of $(\psi, \lambda, \hat{\psi})$, we can find the corresponding value of ϕ by solving for it in Equation 3.1.3. ϕ is called the *zero-score expectation parameter* (ZSEP) because it is defined as the value that makes the expectation of the score function (see Appendix A) evaluated at the point $(\hat{\psi}, \phi)$ equal to zero.

For a given value of $(\psi, \phi, \hat{\psi})$, it is also possible to solve Equation 3.1.3 for λ. This allows us to write Equation 3.1.2 in terms of $L(\psi, \lambda)$:

$$\bar{L}(\psi) = \int_\Phi L(\psi, \lambda(\psi, \phi))\pi(\phi)d\phi. \tag{3.1.4}$$

Note that ϕ depends on the data through $\hat{\psi}$. In most circumstances, such a parameterization would be considered pointless as the reason we collect data in the first place is to estimate parameters. However, from the perspective of the likelihood function, once the data have been collected, they are considered fixed in place, and so there is no issue with using a quantity such as ϕ that depends on the data to parameterize it.

Severini (2018) proved that reparameterizing the nuisance parameter in terms of the ZSEP yields the same desirable properties in the subsequent integrated likelihood when ψ and λ are implicit. Suppose $\psi = \varphi(\theta)$, for some function $\varphi : \Theta \to \Psi$, and consider the set of all values of θ satisfying $\varphi(\theta) = \hat{\psi}$. Call this set $\Omega_{\hat{\psi}}$ so that

$$\Omega_{\hat{\psi}} = \left\{\omega \in \Theta : \varphi(\omega) = \hat{\psi}\right\}. \tag{3.1.5}$$

3.2. Markov Chain Monte Carlo

3.3. The IL Algorithm

CHAPTER 4

Applications

4.1. Multinomial Distribution

Severini (2018) provides an example of the application

4.2. Standardized Mean Difference

References

Basu, Debabrata. 1977. "On the Elimination of Nuisance Parameters." *Journal of the American Statistical Association* 72 (358): 355–66. http://www.jstor.org/stable/2286800.

Berger, James O., Brunero Liseo, and Robert L. Wolpert. 1999. "Integrated Likelihood Methods for Eliminating Nuisance Parameters." *Statistical Science* 14 (1): 1–22. http://www.jstor.org/stable/2676641.

Kalbfleisch, J. D., and D. A. Sprott. 1973. "Marginal and Conditional Likelihoods." *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 35 (3): 311–28. http://www.jstor.org/stable/25049882.

Liseo, Brunero. 1993. "Elimination of Nuisance Parameters with Reference Priors." *Biometrika* 80 (2): 295–304. http://www.jstor.org/stable/2337200.

Severini, Thomas A. 2000. *Likelihood Methods in Statistics*. Oxford University Press.

———. 2007. "Integrated Likelihood Functions for Non-Bayesian Inference." *Biometrika* 94 (3): 529–42. http://www.jstor.org/stable/20441394.

———. 2018. "Integrated Likelihoods for Functions of a Parameter." *Stat* 7 (1): e212. https://doi.org/https://doi.org/10.1002/sta4.212.

———. 2022. "Integrated Likelihood Inference in Multinomial Distributions." *Metron*. https://doi.org/10.1007/s40300-022-00236-x.

APPENDIX A

Chapter 3

A.1. Desirable Properties of the Integrated Likelihood

A.1.1. Property 1

Suppose the likelihood function for a parameter θ can be decomposed as the product $L(\theta) = L_1(\psi)L_2(\lambda)$. Then the integrated likelihood for ψ should satisfy

$$\bar{L}(\psi) = L_1(\psi).$$

A.1.2. Property 2

A.1.3. Property 3

A.1.4. Property 4

A.2. The Score Function

$\ell_\lambda(\psi, \lambda)$ is a sub-vector of the vector-valued function $\ell_\theta(\theta)$ in which each component is equal to the derivative of the log-likelihood with respect to a component of the full model parameter θ. The components of ℓ_λ are the ones from ℓ_θ that correspond to the derivatives with respect to the components of λ.

APPENDIX B

Chapter 4