An Algorithm for Approximating Integrated Likelihood Functions with Applications in Meta-Analysis

By

Timothy Ruel

Faculty Committee:
Dr. Thomas Severini, Department of Statistics and Data Science
Dr. Elizabeth Tipton, Department of Statistics and Data Science
Dr. Wenxin Jiang, Department of Statistics and Data Science

A Doctoral Dissertation Prospectus Submitted to the Department of Statistics and Data Science

Northwestern University
Evanston, Illinois
September 18, 2023

Abstract

Table of Contents

Preface

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.

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.

Background

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."

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.

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.

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.

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. [INSERT EXAMPLE(S) OF NONPARAMETRIC MODELS]

[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.

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. [INSERT EXAMPLE(S) OF SEMIPARAMETRIC MODELS]

In summary, the reader may safely assume that all models to which my research applies 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, I will restrict my attention in my dissertation solely to parametric models.

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. In short, we need the likelihood function.

Let $p(x; \theta)$ denote either the probability density function or the probability mass function for a random variable X with parameter θ, depending on whether X is continuous or discrete, respectively. Suppose we observe that X takes on the value x. Then we define the *likelihood function* for θ as follows:

$$L(\theta) \equiv L(\theta; x) = p(x; \theta), \quad \theta \in \Theta.$$

That is, when our sample consists only of a single observation, the likelihood function for θ is simply equal to the p.d.f. (or p.m.f.) of X.

When X is discrete, this means the likelihood function evaluated at θ may be interpreted as the probability that $X = x$ given that θ is the true parameter value. We forfeit this interpretation when X is continuous as $p(x;\theta)$ no longer represents the probability of an individual point. However,

Note the subtle but crucial distinction in our interpretations of the two quantities, however. While $p(x;\theta)$ is 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.[4] It is crucial to remember that despite our use of probability in its definition a likelihood does not itself represent a probability. For instance, there is no requirement that $\int_\Theta L(\theta;x)d\theta = 1$ as would be expected of a probability density function.

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*.

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

[4]The positioning of the arguments θ and x is a reflection of this difference in perspectives.

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 $\lambda = \sigma^2$.

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.

Pseudolikelihood Functions

Nuisance parameters in a statistical model are a serious hindrance to making inferences about the parameter of interest. Consequently, statisticians have dedicated much time to developing techniques that will eliminate them from their models.

If we let

$$\Theta(\psi) = \{\theta \in \Theta \mid \varphi(\theta) = \psi\},$$

then corresponding to $\psi \in \Psi$ is the set of likelihoods

$$\mathcal{L}_\psi = \{L(\theta) \mid \theta \in \Theta(\psi)\}.$$

Statisticians will often attempt to find a summary of the values in $\mathcal{L}_\psi$ that does not depend on λ. This summary is called a **pseudolikelihood function** and can take several different forms: - Maximizing - Conditioning - Marginalizing* - **Integrating**

The Profile Likelihood.

The Conditional Likelihood.

The Marginal Likelihood.

[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.

The Integrated Likelihood.

The Appeal of the Integrated Likelihood

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

Approximating the Integrated Likelihood

The Zero-Score Expectation Parameter

Markov Chain Monte Carlo

The IL Algorithm

Applications

Multinomial Distribution

Standardized Mean Difference