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

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

Timothy Ruel

The primary focus of this prospectus is to motivate and explain an adapted version of numerical likelihood integration as a method for eliminating nuisance parameters from a statistical model.

Table of Contents

CHAPTER 1

Background

1.1. Introduction

The acquisition of knowledge regarding a population of interest has long been the impetus for the field of statistical inference. In all but the most basic of circumstances, limiting constraints such as time, accessibility, and cost make perfect knowledge of a population essentially impossible to obtain. It therefore becomes necessary to infer properties of the population based on a representative sample of observations drawn from it. The procedures by which these samples might be procured are themselves far from trivial, and indeed an entire branch of statistics has been dedicated to their study. However, for the purposes of this paper we are primarily concerned with what occurs after the sample has been taken, and so we will generally take it for granted that a suitably representative sample of the population already exists.

Suppose $\mathbf{x} = (x_1, ..., x_n)$ is one such sample. What information can we then glean from $\mathbf{x}$ about the population from which it has been drawn? 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? As the questions increase in complexity, so do the techniques required to answer them. Unfortunately the natural chaos of the real world all but guarantees there will never be an instrument capable of completely capturing the intricacies of a population whose properties we wish to infer. Hence, some amount of idealization will always be required in order to proceed.

This idealization typically comes in the form of additional assumptions that we impose on the population of interest with the goal of sacrificing what we hope is only a small amount of accuracy in exchange for a large reduction in complexity. 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

that are capable of producing answers with degrees of accuracy high enough to be useful in their own right. Taken as a whole, they form the basis for what is known a *statistical model*.

The traditional framework for a statistical model begins by assuming that there exists an unknown probability distribution P over the population of interest that generates the data we observe from it. We choose to model this observed data as the realized outcomes of some random variable X that is distributed according to P. We will restrict our attention in this paper to distributions that are either discrete or absolutely continuous, meaning they admit a probability density/frequency function over their support. Let $p(x)$ denote the unknown density/frequency function associated with P and $\mathcal{P}$ the set of functions that we are willing to consider as candidates for $p(x)$. Out of necessity, we will proceed as though our choice of $\mathcal{P}$ always contains $p(x)$ though in reality there is nothing specifically requiring it.

We will also assume the set $\mathcal{P}$ is *parameterized*. That is, we assume there exists a *parameter* θ which indexes $\mathcal{P}$, acting as a label that allows us to differentiate between the densities it contains. For a particular value of the parameter θ, say θ_1, we can refer to its corresponding density in $\mathcal{P}$ with the notation $p(\cdot;\theta)$, and therefore $\mathcal{P}$ itself may be written as $\mathcal{P} = \{p(\cdot;\theta)|\, \theta \in \Theta\}$. Θ is called the *parameter space* and represents the set of all possible values θ can take on.

We can think of θ as acting 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. Making inferences regarding θ is like trying to figure out the particular value to which a population's dial has been set.

In general, a model's parameterization is not unique, and for a given parameter θ, we are free to choose any one-to-one function of θ as a new parameter. Once we have made our choice of parameterization, we will assume that the parameter space Θ does contain a singular true parameter value, which we will denote by θ_0. Note the difference in interpretation between θ_0 and θ. We think of θ_0 as a fixed but unknown constant that represents the value of the parameter corresponding to the true density function in $\mathcal{P}$. Conversely, θ represents an arbitrary parameter value that is allowed to range over all possible elements of Θ, including θ_0.[1]

[1]Note that θ_0 may change over time depending on the population. In such cases, any estimate of θ_0 based on a cross-sectional sample drawn from the population is best thought of as an estimate of the true parameter value during the particular time in which the sample was collected.

Crucially, it must always be possible to identify the parameter in our model on the basis of the data we observe. A model is considered *identifiable* if having perfect knowledge of the population would enable us to determine θ_0 with absolute certainty. This is equivalent to requiring that for some observed data x and any two parameters $\theta_1, \theta_2 \in \Theta$, if $p(x;\theta_1) = p(x;\theta_2)$, 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. For example, suppose Y is distributed uniformly on the interval $(0, \alpha+\beta)$, where $\alpha, \beta > 0$. If we use $\theta = (\alpha, \beta)$ as a parameter for the distribution of Y, then θ is unidentifiable since, for instance, the case where $\theta_1 = (0,1)$ and $\theta_1 = (1,0)$ implies that $p(y;\theta_1) = p(y;\theta_2)$ despite the fact that $\theta_1 \neq \theta_2$. This is obviously an undesirable property for a model to possess, and so we will restrict our attention solely to identifiable models in this paper as a means of avoiding it.

Finally, we must make a choice regarding the dimension of the parameter space Θ when formulating our models. *Parametric* models are defined as having finite-dimensional parameter spaces. Any model that is not parametric is called *nonparametric*. We will restrict our attention in this paper solely to parametric models whose parameter spaces are subsets of $\mathbb{R}^d$, where $d \in \mathbb{Z}^+$.

1.2. The Likelihood Function

Once we have chosen a model, our goal then becomes to identify the true density function in $\mathcal{P}$ or, at the very least, 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 θ_0 itself. 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_0 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.

Given some observed data $X = x$, the *likelihood function* for θ is defined as

$$L(\theta) \equiv L(\theta; x) = p(x;\theta), \;\; \theta \in \Theta. \tag{1.2.1}$$

That is, 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.

When X is discrete, we may interpret $L(\theta;x)$ as the probability that $X = x$ given that θ is the true parameter value. Crucially, this is *not* equivalent to the inverse probability that θ is the true parameter value given $X = x$. The likelihood does not directly tell us anything about the probability that θ assumes any particular value at all. Though intuitively appealing, this interpretation constitutes a fundamental misunderstanding of what a likelihood function is, and great care must be taken to avoid it.

When X is continuous, the likelihood for θ may still be defined as it is in Equation 1.2.1. However, we must 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 however still think of the likelihood as being proportional to the probability that X takes on a value "close" to x, meaning that that X is within a tiny neighborhood of x. 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 when X is continuous to avoid using $L(\theta;x)$ to make probability statements about θ. Despite our use of one in its definition, the likelihood is *not* itself a probability density function for the parameter θ and need not obey the same laws as one.

1.2.1. Transformations

There are a few useful transformations of the likelihood function that we will define here for use in future chapters. The first is the *log-likelihood function*, which is defined as the natural logarithm of the likelihood function:

$$\ell(\theta) \equiv \ell(\theta;\mathbf{x}) = \log L(\theta;\mathbf{x}), \;\; \theta \in \Theta. \tag{1.2.2}$$

In practice, we will typically eschew direct analysis of the likelihood in favor of the log-likelihood due to the nice mathematical properties logarithms possess. Chief among these properties is the ability to turn

products into sums (i.e. $\log(ab) = \log(a) + \log(b)$ for $a, b > 0$). Sums tend to be easier to differentiate than products, making this is a particularly useful feature for likelihood functions, which are often expressed as the product of marginal density functions when the observations are independent.

The other key property of logarithms that makes the log-likelihood so useful is that they are strictly increasing functions of their arguments (i.e. $\log x > \log y$ for $x > y > 0$). This monotonicity ensures that the locations of a function's extrema are preserved when the function is passed to the argument of a logarithm. For example, for a positive function f with a global maximum, $\underset{x}{\operatorname{argmax}} f(x) = \underset{x}{\operatorname{argmax}} \log f(x)$.

We will refer to the derivatives of $\ell(\theta)$ with respect to θ with the notation $\ell_\theta(\theta) = \nabla\ell(\theta)$, $\ell_{\theta\theta}(\theta) = \nabla^2\ell(\theta)$, et cetera. Since we will consider the general case in which θ is a multi-dimensional vector, it follows that $\ell_\theta(\theta)$ is also a vector, $\ell_{\theta\theta}(\theta)$ is a matrix, $\ell_{\theta\theta\theta}(\theta)$ is a three-dimensional array, and so forth.

The first derivative of the log-likelihood with respect to θ appears frequently enough in the analysis of likelihood functions that it has earned its own name - the *score function*. Formally, it is defined as

$$\mathcal{S}(\theta) \equiv \mathcal{S}(\theta; \mathbf{x}) = \ell_\theta(\theta; \mathbf{x}), \ \ \theta \in \Theta. \tag{1.2.3}$$

Similarly, the negative second derivative of the log-likelihood function with respect to θ is called the *observed information*. Formally, it is defined as

$$\mathcal{J}(\theta) \equiv \mathcal{J}(\theta; \mathbf{x}) = -\ell_{\theta\theta}(\theta; \mathbf{x}), \ \ \theta \in \Theta. \tag{1.2.4}$$

The name derives from the fact that it measures the curvature of the log-likelihood around its maximum; the sharper the curve, the less uncertainty we have about θ.

Note that $\ell(\theta; \mathbf{x})$, $\mathcal{S}(\theta; \mathbf{x})$, and $\mathcal{J}(\theta; \mathbf{x})$ are all functions of the data $\mathbf{x}$ and therefore can be interpreted as random variables themselves with respect to $p(\mathbf{x}; \theta)$. Consequently, quantities such as their expectations and variances are well-defined. In particular, the variance of the score function, known as the *expected information* or the *Fisher information*, is another useful transformation that we will consider in more detail in Chapter 2. Formally, it is defined as

$$\mathscr{I}(\theta) = \mathbb{V}\big[\mathcal{S}(\theta; \mathbf{x}); \theta\big], \ \ \theta \in \Theta. \tag{1.2.5}$$

1.2.2. Regularity Conditions

It will be useful to establish some *regularity conditions* for the models under consideration in this paper. The motivations behind the specific choice of conditions can vary, but the primary goal is usually to allow the log-likelihood function and its derivatives to be approximated by a Taylor polynomial of some degree (typically one or two) in θ. For our purposes, we will call a parametric model *regular* if it satisfies the following conditions:

C1) All of the densities in $\mathcal{P}$ have common support which does not depend on θ.

C2) There exists an open subset $\Theta^* \subseteq \Theta$ of which the true parameter value θ_0 is an interior point.

C3) All of the densities in $\mathcal{P}$ are continuously differentiable with respect to θ up to third order for all $\theta^* \in \Theta^*$.

C4) $\mathscr{I}(\theta)$ is positive definite for all $\theta \in \Theta$, and Θ is a convex set.

C5) For each $\theta^* \in \Theta^*$, there exists a neighborhood $N_r(\theta^*) \subseteq \Theta^*$ such that for all $\theta \in N_r(\theta^*)$, the components of the third derivative of the log-likelihood, $\ell_{\theta\theta\theta}(\theta; \mathbf{x})$, are all bounded by a random function $M(\mathbf{x})$ with finite expectation. That is,

$$\sup_{\substack{|\boldsymbol{\alpha}|=3 \\ \boldsymbol{\alpha}\in\mathbb{N}_0^d}} \sup_{\theta \in N_r(\theta^*)} \left| D^{\boldsymbol{\alpha}} \ell(\theta; \mathbf{x}) \right| \leq M(\mathbf{x}) \text{ w.p. } 1$$

where $\mathbb{E}[M(X); \theta_0] < \infty$.

Any model satisfying these conditions will obey the following properties:

P1) Derivatives up to the second order with respect to θ can be passed under the integral sign in $\int dP(\mathbf{x}; \theta)$.

P2) Guaranteed existence, uniqueness, and consistency of the maximum likelihood estimate of θ_0, $\hat{\theta}$.

P3) Convergence of the score function evaluated at θ_0 and $\hat{\theta}$ to multivariate normal distributions as the sample size increases. More precisely,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{MVN}(\mathbf{0}, \mathscr{I}(\theta_0)^{-1}),$$

$$\frac{1}{\sqrt{n}} \mathcal{S}(\theta_0) \xrightarrow{d} \text{MVN}(\mathbf{0}, \mathscr{I}(\theta_0)).$$

1.2.3. 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 $\mathbf{x}$. 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. 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 $L(\theta_1; \mathbf{x}) > L(\theta_2; \mathbf{x})$, then θ_1 is the better estimate of the true parameter value and vice versa.

This can be extended to include as many candidate parameter values as we would like. For n potential estimates of θ_0, 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 Θ. Formally,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta; \mathbf{x}). \tag{1.2.6}$$

The existence and uniqueness of $\hat{\theta}$ is guaranteed for any model satisfying the regularity conditions we have assumed and can be found as the unique solution to the system of equations

$$\mathcal{S}(\theta) = \mathbf{0} \tag{1.2.7}$$

Furthermore, the MLE will be consistent for θ_0 and converge to a normal distribution as the sample size increases.

CHAPTER 2

Pseudolikelihood Analysis

2.1. Model Parameter Decomposition

2.1.1. Introduction

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 different 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*.

Since ψ is of lower dimension than θ, it necessarily follows that there is another parameter λ, taking values in a set $\Lambda \subseteq \mathbb{R}^{k-m}$, that is made up of whatever is "left over" from the full parameter θ. We refer to λ as the *nuisance parameter* due to its ability to complicate inference regarding the parameter of interest. Despite not being the object of study themselves, nuisance parameters are nevertheless capable of modifying the distributions of our observations and therefore must be accounted for when conducting inference or estimation regarding the parameter of interest.[1] The process by which this is accomplished is often nontrivial and can constitute a significant barrier that must be overcome.

While not strictly required, we will assume the parameter of interest ψ is always one-dimensional for the purposes of this paper. That is, $\Psi \subseteq \mathbb{R}$ and consequently $\Lambda \subseteq \mathbb{R}^{k-1}$. This restriction reflects the common habit of researchers to focus on scalar-valued summaries of vector quantities. For example, suppose we observe data $Y = (y_1, ..., y_n)$, where each y_i is the outcome of some random variable $Y_i \sim N(\mu_i, \sigma_i^2)$, and we are interested in estimating the average of the population means, $\frac{1}{n}\sum_{i=1}^{n} \mu_i$. Rather than defining $\psi = (\mu_1, ..., \mu_n)$, we can instead define $\psi = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ directly, bypassing the need to estimate each μ_i individually before taking their average. This does carry the trade-off of increasing the dimension of the nuisance parameter, which must be dealt with before conducting inference or estimation

[1] Note that nuisance parameters are not always uniquely defined. Depending on the choice of parameter of interest, there may be multiple or even infinite ways to define a nuisance parameter.

on ψ. However, as we will discuss in Chapter 3, having a high-dimensional nuisance parameter is not necessarily an issue, especially for the integrated likelihood methods under discussion in this paper which have been shown to work particularly well in situations where the dimension of λ is large relative to the sample size; see, for example, De Bin et al. (2015) and Schumann et al. (2021).

2.1.2. Explicit Parameters

Parameters of interest and nuisance parameters can be broadly classified into two categories, explicit or implicit. For a given statistical model, both types of parameter must occupy the same category - it is not possible for ψ to be explicit and λ to be implicit, or vice versa.

Let us first consider the case in which ψ and λ are *explicit* parameters. This means that 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

$$\psi = (\theta_{I_1}, ..., \theta_{I_m}). \tag{2.1.1}$$

It immediately follows that λ is the sub-vector of all components of θ that are not part of ψ. More precisely, if we let $J = \{J_1, ..., J_{k-m}\} \subsetneq \{1, ..., k\}$ such that $I \cup J = \{1, ..., k\}$ and $I \cap J = \emptyset$, then

$$\lambda = (\theta_{J_1}, ..., \theta_{J_{k-m}}). \tag{2.1.2}$$

θ can therefore be decomposed as $\theta = (\psi, \lambda)$ when ψ and λ are explicit, provided we shuffle the indices appropriately.

2.1.3. Implicit Parameters

Now let us consider the case in which ψ and λ are *implicit* parameters. This means there exists some function $\varphi : \Theta \to \Psi$ for which the parameter of interest can be written as

$$\psi = \varphi(\theta). \tag{2.1.3}$$

As before, 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 again implies the existence of a nuisance parameter $\lambda \in \Lambda \subseteq \mathbb{R}^{k-m}$. However, unlike in the explicit case, a closed form expression for λ in terms of the original components of θ need not exist. For this reason, implicit nuisance parameters are in general more difficult to eliminate compared to their explicit counterparts.

Note that when the parameter of interest and nuisance parameter are explicit, it is always possible to define a function φ such that

$$\varphi(\theta) = (\theta_{I_1}, ..., \theta_{I_m}) \equiv \psi, \tag{2.1.4}$$

where $\{I_1, ..., I_m\}$ is defined as above. Hence, the first case is really just a special example of this more general one in which $\psi = \varphi(\theta)$. With this understanding in mind, we will use the notation $\psi = \varphi(\theta)$ to refer to the parameter of interest in general, only making the distinction between implicitness and explicitness when the difference is relevant to the situation.

2.2. Pseudolikelihood Functions

2.2.1. Introduction

The natural solution to the hindrance nuisance parameters pose to making inferences on the parameter of interest is to find a method for eliminating them from the model altogether. Since one way of uniquely specifying a model is through its likelihood function, this is equivalent to eliminating the nuisance parameters from the likelihood function itself. The result of this elimination is what is known as a pseudolikelihood function.

In general, a *pseudolikelihood function* for ψ is defined as being a function of ψ and the data alone, having properties resembling that of a genuine likelihood function. Suppose $\psi = \varphi(\theta)$ for some function ϕ and parameter $\theta \in \Theta$. If we let $\Theta(\psi) = \{\theta \in \Theta \ : \ \varphi(\theta) = \psi\}$, then associated with each $\psi \in \Psi$ is the set of likelihoods $\mathcal{L}_\psi = \{L(\theta) \ : \ \theta \in \Theta(\psi)\}$.

Any summary of the values in $\mathcal{L}_\psi$ that does not depend on λ theoretically constitutes a pseudolikehood function for ψ. 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.

2.2.2. The Profile Likelihood

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

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 $\lambda = \sigma^2$.

2.2.3. The Conditional Likelihood

2.2.4. The Marginal Likelihood

2.2.5. The Integrated Likelihood

2.3. Asymptotic Analysis of Likelihoods and Pseudolikelihoods

Knowledge of how likelihood functions behave as the sample size increases is a useful tool for understanding the properties of the estimates they produce. We will base our analysis of this asymptotic behavior on second-order Taylor expansions of the log-likelihood function and its derivatives.

2.3.1. The Bartlett Identities

The Bartlett identities are a set of equations relating to the expectations of functions of derivatives of a log-likelihood function. A well-specified genuine likelihood function will automatically satisfy each of the Bartlett identites; however, an arbitrary function of θ and X will not. For this reason, the identities act as a litmus test of sorts for determining the validity of a pseudolikelihood as an approximation to the genuine likelihood from which it originated.[2]

[2]The Bartlett identities offer an alternative way of characterizing the difference between likelihood and pseudolikelihood functions. A genuine likelihood function of θ is any nonnegative random function of θ for which all of the Bartlett identities hold. A pseudolikelihood of θ is any nonnegative random function of θ for which at least one of the Bartlett identities does not hold.

Consider the case in which a random variable X has a probability density f that depends on a scalar parameter θ. Denote the log-likelihood function for θ by $\ell(\theta; x) = \log f(x; \theta)$ and its first derivative with respect to θ by $\ell_\theta(\theta; x) = \frac{\partial}{\partial\theta}\ell(\theta; x)$. We previously assumed in Section 2.1.4. that all probability distributions for which the results in this paper apply are regular. One consequence of this assumption is that derivatives and integrals of the density functions for these distributions may be interchanged. Now, taking the expectation of $\ell_\theta(\theta; x)$ gives

$$\begin{aligned}
\mathbb{E}[\ell_\theta(\theta; x); \theta] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta; x); \theta\right] \\
&= \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta}\ell(\theta; x)\right] f(x; \theta)dx \\
&= \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta}\log f(x; \theta)\right] f(x; \theta)dx \\
&= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta}f(x; \theta)}{f(x; \theta)} f(x; \theta)dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta}f(x; \theta)dx \\
&= \frac{d}{d\theta}\int_{\mathbb{R}} f(x; \theta)dx \qquad \text{(by regularity of } f\text{)} \\
&= \frac{d}{d\theta}1 \\
&= 0.
\end{aligned}$$

Therefore,

$$\mathbb{E}[\ell_\theta(\theta; x); \theta] = 0 \text{ for all } \theta. \tag{2.3.1}$$

Equation 2.3.1 is called the first Bartlett identity. In words, it states that the expectation of the first derivative of the log-likelihood function of a statistical model with respect to the model parameter will always be 0. Another name for ℓ_θ is the *score function*, and any pseudolikelihood that also satisfies the first Bartlett identity is said to be *score-unbiased*.

If we now consider the second derivative of $\ell(\theta; x)$, we have

$$
\begin{aligned}
\ell_{\theta\theta}(\theta;x) &= \frac{\partial^2}{\partial\theta^2}\ell(\theta;x) \\
&= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ell(\theta;x)\right] \\
&= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] \\
&= \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right] \\
&= \frac{\left[\frac{\partial^2}{\partial\theta^2}f(x;\theta)\right]f(x;\theta) - \left[\frac{\partial}{\partial\theta}f(x;\theta)\right]\left[\frac{\partial}{\partial\theta}f(x;\theta)\right]}{\left[f(x;\theta)\right]^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left[\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left[\ell_\theta(\theta;x)\right]^2.
\end{aligned}
$$

Rearranging terms and taking expectations yields

$$
\begin{aligned}
\mathbb{E}[\ell_{\theta\theta}(\theta;x);\theta] + \mathbb{E}\Big[\left(\ell_\theta(\theta;x)\right)^2;\theta\Big] &= \mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)};\theta\right] \\
&= \int_{\mathbb{R}}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)}\right]f(x;\theta)dx \\
&= \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx \\
&= \frac{d^2}{d\theta^2}\int_{\mathbb{R}}f(x;\theta)dx \qquad \text{(by regularity of } f\text{)} \\
&= \frac{d^2}{d\theta^2}1 \\
&= 0.
\end{aligned}
$$

Therefore,

$$
\mathbb{E}[\ell_{\theta\theta}(\theta;x);\theta] + \mathbb{E}\Big[\left(\ell_\theta(\theta;x)\right)^2;\theta\Big] = 0 \text{ for all } \theta. \tag{2.3.2}
$$

Equation 2.3.2 is called the second Bartlett identity. The second term on the left-hand side can be further rewritten as

$$\begin{aligned}\mathbb{E}\Big[\big(\ell_\theta(\theta;x)\big)^2;\theta\Big] &= \mathbb{V}[\ell_\theta(\theta;x);\theta] + \Big(\mathbb{E}[\ell_\theta(\theta;x);\theta]\Big)^2 \\ &= \mathbb{V}[\ell_\theta(\theta;x);\theta]. \qquad \text{(by the first Bartlett identity)}\end{aligned}$$

Another name for this quantity is the *expected information*. It follows from the second Bartlett identity that

$$\mathbb{E}[-\ell_{\theta\theta}(\theta;x);\theta] = \mathbb{V}[\ell_\theta(\theta;x);\theta]. \tag{2.3.3}$$

The quantity $-\ell_{\theta\theta}(\theta;x)$ is called the *observed information*. Any pseudolikelihood that satisfies the second Bartlett identity is said to be *information-unbiased*.

It is possible to derive further Bartlett identities by continuing in this manner for an arbitrary number of derivatives of the log-likelihood function, provided that they exist. However, the first two are sufficient for our purposes of evaluating the validity of pseudolikelihoods as approximations to a genuine likelihood so we will not go further here. While the above derivations were performed under the assumption that θ is a scalar, the Bartlett identities also hold in the case where θ is a multi-dimensional vector.

2.3.2. Single-Index Asymptotic Theory

Single-index asymptotic theory describes the behavior of likelihood-based statistics as the sample size (n) grows to infinity while the dimension of the nuisance parameter (m) remains fixed. The aim of this section is to present a basic overview of the theory's results so that the reader will have a readily available baseline against which to compare the results of the following section discussing two-index asymptotic theory, in which the dimension of the nuisance parameter is allowed to increase with the sample size.

We will couch our analysis in the framework of independent and identically distributed (IID) observations in order to demonstrate a standard scenario in which likelihood theory holds. This framework is by no means the only one to which the later results of this paper apply, and in fact there are plenty of models without IID observations that produce likelihood-based estimates with similar properties to those

from the IID case (e.g. asymptotic normality). However, understanding how the theory works in the IID case makes it relatively straightforward to extend the methods to these other cases.

Let $X_1, ..., X_n$ be IID random variables with density function $p(\mathbf{x}; \theta)$ and let $\hat{\theta}$ denote the maximum likelihood estimate for θ_0 based on the observed values $x_1, ..., x_n$. The traditional method for analyzing the asymptotic behavior of $\hat{\theta}$ is to use a first-order Taylor series expansion of the score function.[3]

To do this, first note that the likelihood function for θ based on $\mathbf{x} = (x_1, ..., x_n)$ is

$$L(\theta; \mathbf{x}) \equiv p(\mathbf{x}; \theta) = \prod_{i=1}^{n} p(x_i; \theta).$$

We may therefore write the log-likelihood function as

$$\begin{aligned} \ell(\theta; \mathbf{x}) &= \log L(\theta; \mathbf{x}) \\ &= \log p(\mathbf{x}; \theta) \\ &= \log \left[\prod_{i=1}^{n} p(x_i; \theta) \right] \\ &= \sum_{i=1}^{n} \log p(x_i; \theta) \\ &= \sum_{i=1}^{n} \ell(\theta; x_i). \end{aligned}$$

It follows that the score function can be written as

$$\begin{aligned} \mathcal{S}(\theta; \mathbf{x}) &= \frac{\partial}{\partial \theta} \ell(\theta; \mathbf{x}) \\ &= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ell(\theta; x_i) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ell(\theta; x_i) \\ &= \sum_{i=1}^{n} \mathcal{S}(\theta; x_i). \end{aligned} \tag{2.3.4}$$

[3]See Appendix A for a review of Taylor's theorem and the conditions under which it is satisfied.

In other words, the score function for the parameter θ based on data $x_1, ..., x_n$ can be written as the sum of individual contributions $\mathcal{S}(\theta; x_i)$, $i = 1, ..., n$, where each $\mathcal{S}(\theta; x_i)$ is the score function for θ based only on observation x_i. Note that these individual contributions are all independent from one another, a consequence of our earlier assumption that the observations themselves are independent.

Equation 2.3.4 implies that a Taylor series expansion of $\mathcal{S}(\theta; \mathbf{x})$ will be equal to the sum of the Taylor series expansions of its individual contributions, plus a remainder term that grows with n. We have assumed each x_i is identically distributed, so it suffices to consider the expansion for an arbitrary contribution, $\mathcal{S}(\theta; x_i)$.

Since θ is a vector, the k-th derivative of the log-likelihood with respect to θ will be a k-dimensional array having d entries along each of its k indices. In particular, the score function (i.e. the first derivative of the log-likelihood) will be a $d \times 1$ vector,

$$\mathcal{S}(\theta; x_i) = \begin{pmatrix} \mathcal{S}_1(\theta; x_i) \\ \vdots \\ \mathcal{S}_d(\theta; x_i) \end{pmatrix},$$

where each component is a function $S_j : \Theta \to \mathbb{R}$. Similarly, the first and second derivatives of the score function will be a $d \times d$ matrix and a three-dimensional $d \times d \times d$ array, respectively. To simplify notation, we will perform a componentwise Taylor series expansion of the score function around the point $\theta = \theta_0$ wherein each component $S_j(\theta; x_i)$ of $S(\theta; x_i)$ is expanded separately.

Assume that the regularity conditions established in Section 1.2.2 apply to our model.[4] Conditions C2) and C3) stipulate that there exists an open subset $\Theta^* \subseteq \Theta$ containing θ_0 for which $S_j(\theta; x_i)$ is twice continuously differentiable with respect to θ for all $\theta \in \Theta^*$. C5) requires that the components of the second derivative of the score function are all bounded (with probability 1) by a random function $M(\mathbf{x})$ with finite expectation for all $N_r(\theta_0)$. Together, these conditions satisfy the necessary requirements to apply Taylor's theorem with the Lagrange form of the remainder term to the score function. Hence, for

[4]Since it is the score function that we are expanding, we will phrase our references to the regularity conditions from Section 1.2.2 with respect to it, without making direct mention of the density or log-likelihood functions in terms of which the conditions were originally framed. For example, referring to the third derivative of the log-likelihood is equivalent to referring to the second derivative of the score; we will use only the latter phrasing and not the former in this section.

all $\theta \in N_r(\theta_0)$, the first order Taylor series expansion of $S_j(\theta; x_i)$ around θ_0 is given by

$$T_j(\theta; x_i) = S_j(\theta_0; x_i) + \nabla S_j(\theta_0; x_i)^T(\theta - \theta_0) + R(\theta),$$

where the remainder term $R(\theta)$ satisfies

$$|R(\theta)| \leq \frac{M(\mathbf{x})}{\boldsymbol{\alpha}!}|\theta - \theta_0|^{\boldsymbol{\alpha}}$$

for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = 2$.

2.3.3. Two-Index Asymptotic Theory

Two-index asymptotic theory describes the behavior of likelihood functions as the sample size and dimension of the nuisance parameter both tend to infinity, with m growing at least as fast as n. De Bin (2015)

CHAPTER 3

Approximating the Integrated Likelihood Function

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 B) 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 constant 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* (ZSE) parameter because it is defined as the value that makes the expectation of the score function (where the derivative is taken with respect to λ) evaluated at the point $(\hat{\psi}, \phi)$ equal to zero.

Note that ϕ is a function of the data through $\hat{\psi}$. Normally we avoid creating such dependencies in our parameters as it renders them useless for the purpose of parameterizing a statistical model. However, from the perspective of the likelihood function, once the data have been collected they are considered fixed in place and there is no issue with using a quantity such as ϕ that depends on the data to parameterize it.

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 $L(\psi, \lambda)$:

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

Severini (2018) proved that reparameterizing the nuisance parameter in terms of the ZSE parameter 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

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

Elements of $\Omega_{\hat{\psi}}$ take the form $(\hat{\psi}, \phi)$, where $\phi \in \Lambda$.

3.1.1. Weight Functions

CHAPTER 4

Applications

4.1. Multinomial Distribution

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.

De Bin, Riccardo, Nicola Sartori, and Thomas A. Severini. 2015. "Integrated likelihoods in models with stratum nuisance parameters." *Electronic Journal of Statistics* 9 (1): 1474–91. https://doi.org/10.1214/15-EJS1045.

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.

Schumann, Martin, Thomas A. Severini, and Gautam Tripathi. 2021. "Integrated Likelihood Based Inference for Nonlinear Panel Data Models with Unobserved Effects." *Journal of Econometrics* 223 (1): 73–95. https://doi.org/10.1016/j.jeconom.2020.10.001.

———. 2023. "The Role of Score and Information Bias in Panel Data Likelihoods." *Journal of Econometrics* 235 (2): 1215–38. https://doi.org/10.1016/j.jeconom.2022.08.011.

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/10.1002/sta4.212.

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

APPENDIX A

Chapter 2

A.1. Definitions and Notation

A.1.1. Neighborhoods

A *neighborhood* of a point $\mathbf{p} \in \mathbb{R}^d$ is the set of all points $\mathbf{x} \in \mathbb{R}^d$ such that the Euclidean distance between $\mathbf{p}$ and $\mathbf{x}$ is less than some radius $r > 0$. We use the following notation to refer to such neighborhoods:

$$N_r(\mathbf{p}) = \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x} - \mathbf{p}|| < r\}.$$

A.1.2. Line Segments

For two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, a third point $\bar{\mathbf{x}}$ is said to be on the *line segment* connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ if there exists $\omega \in [0, 1]$ such that $\bar{\mathbf{x}} = \omega\mathbf{x}_1 + (1 - \omega)\mathbf{x}_2$. We use the following notation to refer to such line segments:

$$LS(\mathbf{x}_1, \mathbf{x}_2) = \{\omega\mathbf{x}_1 + (1 - \omega)\mathbf{x}_2 : \omega \in [0, 1]\}.$$

A.1.3. Higher Order Partial Derivatives

For $\boldsymbol{\alpha} \in \mathbb{N}^d$ and $\mathbf{x} \in \mathbb{R}^d$, we define the following multi-index notation:

$$|\boldsymbol{\alpha}| = \alpha_1 + ... + \alpha_d, \;\; \boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_d!, \;\; \mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}.$$

For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is k times differentiable at the point $\mathbf{a} \in \mathbb{R}^d$, define the following notation to indicate the mixed partial derivatives of f at $\mathbf{a}$ up to and including order k:

$$D^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}, \;\; |\boldsymbol{\alpha}| \leq k.$$

A.2. Taylor's Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that is $(k+1)$-times continuously differentiable in a neighborhood $N_r(\mathbf{x}_0)$ of some point $\mathbf{x}_0 \in \mathbb{R}^d$ and suppose there exists M satisfying $|D^{\boldsymbol{\alpha}} f| \leq M$ for all $x \in N_r(x_0)$ and all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = k+1$. Then

$$f(\mathbf{x}) = \sum_{0 \leq |\boldsymbol{\alpha}| \leq k} \frac{D^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}} + R_k(\mathbf{x}),$$

where the remainder term $R_k(\mathbf{x})$ satisfies

$$|R_k(\mathbf{x})| \leq \frac{M}{\boldsymbol{\alpha}!} |\mathbf{x} - \mathbf{x}_0|^{\boldsymbol{\alpha}}$$

for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = k+1$

APPENDIX B

Chapter 3

B.1. Desirable Properties of the Integrated Likelihood

B.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

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

B.1.2. Property 2

B.1.3. Property 3

B.1.4. Property 4

APPENDIX C

Chapter 4