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

Table of Contents

CHAPTER 1

Background

1.1. 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 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 that are capable of producing answers with degrees of accuracy high enough as to be useful in their own right.

1.1.1. Data-Generating Mechanism

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 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 $\mathcal{S}$. Out of necessity, we will assume that $\mathcal{P}$ always contains this distribution though in reality there is nothing specifically requiring it.

1.1.2. Model Parameterization

Another assumption found in almost every model is that the set $\mathcal{P}$ is considered to be *parameterized*. That is, we assume there exists a *parameter* θ that indexes $\mathcal{P}$, acting as a label that allows us to differentiate between the distributions it contains. For a particular value of the parameter θ, say θ_1, we can refer to its corresponding distribution in $\mathcal{P}$ with the notation $\mathcal{P}_{\theta_1}$, and therefore $\mathcal{P}$ itself may be written as $\mathcal{P} = \{\mathcal{P}_\theta | \, \theta \in \Theta\}$. Θ is called the *parameter space* and represents the set of all possible values θ can take on.

We can think of the parameter 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. Much of statistical inference can be boiled down to figuring out the particular value to which a population's dial has been set. For a given population, we will assume that a true parameter value does always exist and is contained in Θ. Let the symbol θ_0 denote this value.[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.

In general, a model's parameterization is not unique, and for any parameter θ, we may choose any one-to-one function of θ as a new parameter.

1.1.3. Identifiability

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.[2] 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

[1] The difference between θ_0 and θ is subtle but important. θ_0 is a fixed but unknown constant that represents the value of the parameter corresponding to the "true" distribution in $\mathcal{P}$. θ represents an arbitrary element taken from the space of all possible parameter values, Θ.

[2] 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.

probability distribution. For example, suppose X is distributed uniformly on the interval $(0, \alpha + \beta)$, for two numbers $\alpha, \beta > 0$. If we use $\theta = (\alpha, \beta)$ as a parameter for the distribution of X, then θ is unidentifiable since, for instance, if $\theta_1 = (0, 1)$ and $\theta_1 = (1, 0)$ then $\mathcal{P}_{\theta_1} = \mathcal{P}_{\theta_2}$ 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.

1.1.4. 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. When $\Theta \subseteq V$, where V is an infinite-dimensional space, the model is said to be a *nonparametric*. 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.

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.

Models for which Θ is of finite dimension are called *parametric*. That is, $\Theta \subseteq \mathbb{R}^k$, where $k \in \mathbb{Z}^+$. This is the most common type of model used by statisticians, with examples including the normal family of distributions as well as the Poisson family. For the purposes of this paper, we will assume all statistical models under discussion are parametric in form.

1.2. 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.

1.2.1. The Discrete Case

Suppose X is a discrete random variable with frequency function $p(x;\theta)$. For a given sample $X = x$, 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, 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. Crucially, 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.

1.2.2. The Continuous Case

When X is instead a continuous random variable, the likelihood for θ may still be defined as it is in Equation 1.2.1. However, $p(x;\theta)$ 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 however still think of the likelihood as being proportional to the probability that X is "close" to the value x.[3] 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

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

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.

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 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 $\widehat{\theta}$ (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,

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta; x). \tag{1.2.2}$$

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

$$\ell(\theta) \equiv \ell(\theta; x) = \log L(\theta; x), \ \ \theta \in \Theta. \tag{1.2.3}$$

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 $L(\theta)$.

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

2.3.1. Regularity Conditions

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. In order to guarantee that these expansions exist, we will consider only a particular class of statistical models that obey certain *regularity conditions*. The exact specifications of these conditions can vary depending on the requirements of the researcher, but they are usually made with the goal of ensuring that the log-likelihood function obeys various "nice" properties. Typical examples of these properties include asymptotic normality of the log-likelihood and its first derivative as well as guaranteed existence and consistency of the MLE.

Suppose we have a model with a family of distributions $\mathcal{P}$ indexed by a parameter $\theta \in \Theta$. Let θ_0 denote the true value of θ. For our purposes, a parametric model is called *regular* if it satisfies the following conditions:

1) All of the distributions in $\mathcal{P}$ are absolutely continuous with respect to a σ-finite measure μ and therefore admit a density function $p(x;\theta)$;
2) Any observations $X_1, ..., X_n$ that we draw from $\mathcal{S}$ are independent and identically distributed according to $p(x;\theta)$;
3) The densities
4) The parameter space Θ contains an open subset Θ_0 of which the true parameter value θ_0 is an interior point.

All of the distributions in $\mathcal{P}$ have common support; 2) All of the distributions in $\mathcal{P}$ are absolutely continuous with respect to a sigma-finite measure μ; 3) Any observations $X_1, ..., X_n$ taken from $\mathcal{S}$ are independent and identically distributed with probability density function $f(x;\theta)$ with respect to μ.

2.3.2. 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]

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

[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.

that derivatives and integrals of the density functions for these distributions may be interchanged. Now, taking the expectation of $\ell_\theta(\theta; x)$ gives

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

Therefore,

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

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. Note that 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.

2.3.3. Single-Index Asymptotic Theory

Single-index asymptotic theory describes the behavior of pseudolikelihood functions as the sample size (n) grows to infinity while the dimension of the nuisance parameter (m) remains fixed.

Let $X_1, ..., X_n$ be independent and identically distributed observations taken from some random variable X with distribution P_θ and parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Since we have assumed $\mathcal{P}_\theta$ is absolutely

Let $\hat{\theta}_n$ denote the maximum likelihood estimate for θ.

A second-order Taylor expansion of the log-likelihood $\ell(\theta;x) = \log f(x;\theta)$ around the point $\theta = \hat{\theta}_n$ is given as follows:

$$\ell(\theta) = \ell(\hat{\theta}_n) + \ell_\theta(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \frac{1}{2}\ell_{\theta\theta}(\hat{\theta}_n)(\theta - \hat{\theta}_n)^2 + R_n(\theta),$$

where $R_n(\theta) = \frac{1}{6}\ell_{\theta\theta\theta}(\theta^*)(\theta - \hat{\theta}_n)^3$, for some $\theta^* \in \Theta$.

A second-order Taylor expansion of the score function $\ell_\theta(\theta)$ around the point $\theta = \hat{\theta}_n$ is given as follows:

$$\ell_\theta(\theta) = \ell_\theta(\hat{\theta}_n) + \ell_{\theta\theta}(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \frac{1}{2}\ell_{\theta\theta\theta}(\hat{\theta}_n)(\theta - \hat{\theta}_n)^2 + R_n(\theta),$$

where $R_n(\theta) = \frac{1}{6}\ell_{\theta\theta\theta\theta}(\theta^*)(\theta - \hat{\theta}_n)^3$, for some $\theta^* \in \Theta$.

2.3.4. Two-Index Asymptotic Theory

Two-index asymptotic theory describes the behavior of pseudolikelihood functions as n and m both tend to infinity, with m growing at least as fast as n.

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 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 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 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

APPENDIX B

Chapter 4