Integrated Likelihood Inference in Poisson Distributions

Timothy Ruel

Introduction

Consider a vector $\theta = (\theta_1, ..., \theta_n)$ in which each component represents the mean of a distinct Poisson process. The purpose of this paper is to discuss the task of conducting likelihood-based inference for a real-valued parameter of interest $\psi = \tau(\theta)$. In particular, we will examine the utility of the integrated likelihood function as a tool for obtaining interval and point estimates for ψ, using the performance of the more easily calculated profile likelihood as a benchmark.

We may obtain a sample of values from each Poisson process through repeated measurements of the number of events it generates over a fixed period of time. Suppose we have done so, and let X_{ij} represent the jth count from the ith sample, so that $X_{ij} \sim \text{Poisson}(\theta_i)$ for $i = 1, ..., n$ and $j = 1, ..., m_i$. The probability mass function (pmf) for a single observation $X_{ij} = x_{ij}$ is

$$p(x_{ij};\, \theta_i) = \frac{e^{-\theta_i}\theta_i^{x_{ij}}}{x_{ij}!},\;\; x_{ij} = 0, 1, 2, ...;\;\; \theta_i > 0. \tag{0.1}$$

Denote the sample of counts from the ith process by the vector $X_{i\bullet} = (X_{i1}, ..., X_{im_i})$, its associated mean by $\bar{X}_{i\bullet} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}$, and assume that all of the counts both within and between samples are measured independently. The likelihood function for an individual component θ_i based on the data $X_{i\bullet} = x_{i\bullet}$ is then equal to the product of the individual probabilities of the observed counts, i.e.

$$\begin{aligned}
L(\theta_i; x_{i\bullet}) &= \prod_{j=1}^{m_i} p(x_{ij}; \theta_i) \\
&= \prod_{j=1}^{m_i} \frac{e^{-\theta_i}\theta_i^{x_{ij}}}{x_{ij}!} \\
&= \left(\prod_{j=1}^{m_i} e^{-\theta_i}\right)\left(\prod_{j=1}^{m_i} \theta_i^{x_{ij}}\right)\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1} \\
&= \left(e^{-\sum_{j=1}^{m_i}\theta_i}\right)\left(\theta_i^{\sum_{j=1}^{m_i} x_{ij}}\right)\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1} \\
&= e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1}.
\end{aligned} \tag{0.2}$$

Since L is only useful to the extent that it informs our understanding of the value of θ_i, we are free to replace it with any other function differing from it by just a (nonzero) multiplicative term that is constant with respect to θ_i, provided that the result still satisfies the necessary regularity conditions, as this will not change any conclusions regarding θ_i that we draw from it. Hence, we may safely discard the term in parentheses on the final line of Equation 0.2 as it does not depend on θ_i and instead simply write

$$L(\theta_i; x_{i\bullet}) = e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}. \tag{0.3}$$

It will generally be more convenient to work with the log-likelihood function, which is given by

$$\begin{aligned}
\ell(\theta_i; x_{i\bullet}) &= \log L(\theta_i; x_{i\bullet}) \\
&= \log\left(e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}\right) \\
&= -m_i\theta_i + m_i\bar{x}_{i\bullet}\log\theta_i \\
&= m_i(\bar{x}_{i\bullet}\log\theta_i - \theta_i).
\end{aligned} \tag{0.4}$$

The sum of the log-likelihood functions for each component of θ then forms the basis of the log-likelihood function for θ itself:

$$\begin{aligned}\ell(\theta; x_{1\bullet}, ..., x_{n\bullet}) &= \log L(\theta; x_{1\bullet}, ..., x_{n\bullet}) \\ &= \log\left(\prod_{i=1}^{n} L(\theta_i; x_{i\bullet})\right) \\ &= \sum_{i=1}^{n} \log L(\theta_i; x_{i\bullet}) \\ &= \sum_{i=1}^{n} \ell(\theta_i; x_{i\bullet}) \\ &= \sum_{i=1}^{n} m_i(\bar{x}_{i\bullet} \log\theta_i - \theta_i).\end{aligned} \tag{0.5}$$

We can derive the maximum likelihood estimate (MLE) for θ_i by differentiating Equation 0.4 with respect to θ_i, setting the result equal to 0, and solving for θ_i. This gives the nice result that the MLE is simply equal to the mean of the sample of data $X_{i\bullet}$. That is,

$$\hat{\theta}_i = \bar{X}_{i\bullet}. \tag{0.6}$$

Similarly, the MLE for the full parameter θ is just the vector of MLEs for its individual components:

$$\hat{\theta} \equiv (\hat{\theta}_1, ..., \hat{\theta}_n) = (\bar{X}_{1\bullet}, ..., \bar{X}_{n\bullet}). \tag{0.7}$$

Pseudolikelihoods

Let $\Theta \subseteq \mathbb{R}^n_+$ represent the space of possible values for θ and suppose we have a real-valued *parameter of interest* $\psi = \tau(\theta)$, where $\tau : \Theta \to \Psi$ is a known function with at least two continuous derivatives. Though it is not strictly necessary, in order to align with the tendency of researchers to focus on one-dimensional summaries of vector quantities we will assume for our purposes that ψ is a scalar, i.e. $\Psi \subseteq \mathbb{R}$.

This reduced dimension of Ψ relative to Θ implies the existence of a *nuisance parameter* $\lambda \in \Lambda \subseteq \mathbb{R}^{n-1}$. As its name suggests, λ tends to obfuscate or outright preclude inference regarding ψ and typically must be eliminated from the likelihood before proceeding. The product of this elimination is called a *pseudolikelihood function*. Any function of the data and ψ alone could theoretically be considered a pseudolikelihood, though course in practice some are more useful than others. If we let $\Theta_\psi = \{\theta \in \Theta : \tau(\theta) = \psi\}$, then associated with each $\psi \in \Psi$ is the set of likelihood values $\mathcal{L}_\psi = \{L(\theta) : \theta \in \Theta_\psi\}$. For a given value of ψ, there may exist multiple corresponding values of λ.

We can construct pseudolikelihoods for ψ through clever choices by which to summarize $\mathcal{L}_\psi$ over all possible values of λ. Among the most popular methods of summary are profiling (i.e. maximization), conditioning, and integration, each with respect to the nuisance parameter. These summaries do come at a cost, however; eliminating a model's nuisance parameter from its likelihood almost always sacrifices some information about its parameter of interest as well. One measure of a good pseudolikelihood, therefore, is the balance it strikes between the amount of information it retains about ψ and the ease with which it can be computed.

The Profile Likelihood

The most straightforward method we can use to construct a pseudolikelihood (or equivalently, a pseudo-log-likelihood) function for ψ is usually to find the maximum of $\ell(\theta)$ over all possible of values of θ for each value of ψ. This yields what is known as the *profile* log-likelihood function, formally defined as

$$\ell_p(\psi) = \sup_{\theta \in \Theta \colon \tau(\theta) = \psi} \ell(\theta), \;\; \psi \in \Psi. \tag{0.8}$$

In the case where an explicit nuisance parameter λ exists so that θ may be written as $\theta = (\psi, \lambda)$, Equation 0.8 is equivalent to replacing λ with $\hat{\lambda}_\psi$, its conditional MLE given ψ:

$$\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi). \tag{0.9}$$

Historically, the efficiency with which the profile is capable of producing accurate estimates of ψ relative to its ease of computation has made it the method of choice for statisticians when performing likelihood-based inference regarding a parameter of interest. Examples of profile-based statistics are the MLE for ψ, i.e.,

$$\hat{\psi} = \arg\sup_{\psi \in \Psi} \ell_p(\psi), \tag{0.10}$$

and the signed likelihood ratio statistic for ψ, given by

$$R_\psi = \operatorname{sgn}(\hat{\psi} - \psi)(2(\ell_p(\hat{\psi}) - \ell_p(\psi)))^{\frac{1}{2}}. \tag{0.11}$$

The Integrated Likelihood

The *integrated likelihood* for ψ seeks to summarize $\mathcal{L}_\psi$ by its average value with respect to some weight function π over the space Θ_ψ. From a theoretical standpoint, this is preferable to the maximization procedure found in the profile likelihood as it naturally incorporates our uncertainty regarding the nuisance parameter's true value into the resulting pseudolikelihood. The general form of an integrated likelihood function is given

$$\bar{L}(\psi) = \int_{\Theta_\psi} L(\theta)\pi(\theta; \psi) d\theta. \tag{0.12}$$

It is up to the researcher to choose the weight function $\pi(\cdot; \psi)$, which plays an important role in the properties of the resulting integrated likelihood. Severini (2007) developed a method for re-parameterizing λ that makes the integrated likelihood relatively insensitive to the exact weight function chosen. Using this new parameterization, we have great flexibility in choosing our weight function; as long as it does not depend on the parameter of interest, the integrated likelihood that is produced will enjoy many desirable frequency properties.

Application to Poisson Models

We now turn our attention to the task of using the ZSE parameterization to construct an integrated likelihood that can be used to make inferences regarding a parameter of interest derived from the Poisson model described in the introduction. We will

Estimating the Weighted Sum of Poisson Means

Consider the weighted sum

$$Y = \sum_{i=1}^{n} w_i X_i, \tag{0.13}$$

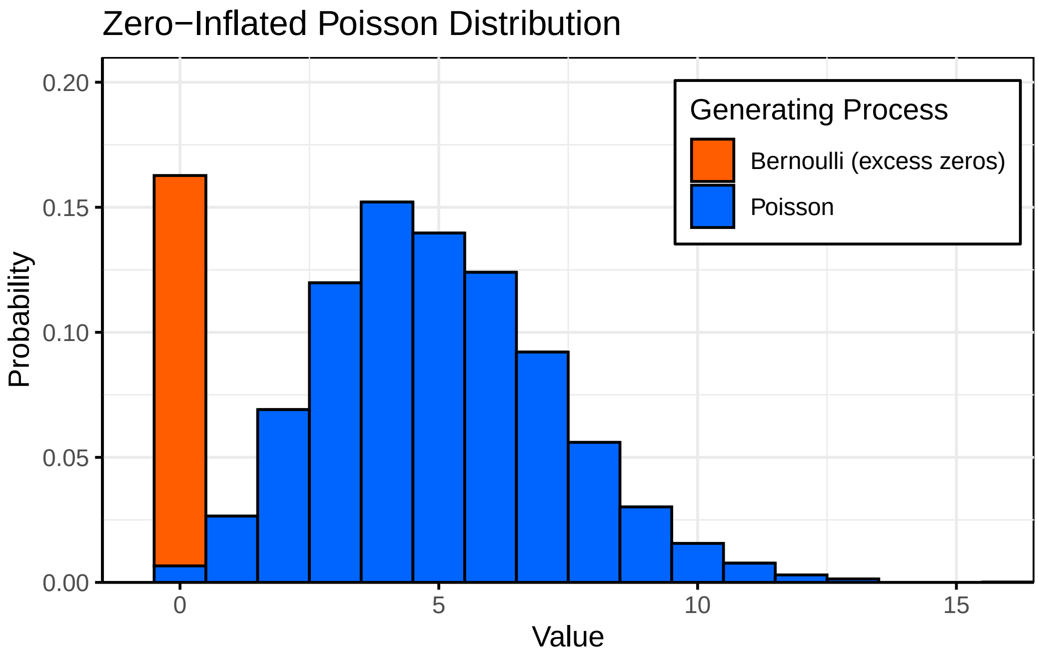
where each w_i is a known constant greater than zero. Suppose we take for our parameter of interest the expected value of this weighted sum, so that

$$\psi \equiv \mathrm{E}(Y) = \sum_{i=1}^{n} w_i \theta_i. \tag{0.14}$$

Examples

Zero-Inflated Poisson Regression

A sample of count data is called *zero-inflated* when it contains an excess amount of zero-valued observations. A common tactic to account for this excess is to model the data using a mixture of two processes, one that generates zeros and another that generates counts, some of which may also be zeros. When this count-generating process follows a Poisson distribution, we call the resulting mixture a zero-inflated Poisson (ZIP) model.



Let $U \sim \text{Bernoulli}(1-\pi)$ and $V \sim \text{Poisson}(\mu)$ Suppose U and V are independent and let $Y = UV$. Then $Y \sim \text{ZIP}(\mu, \pi)$. To derive its distribution, we begin by recognizing that $Y = 0$ when either $U = 0$ or $V = 0$ so that

$$\begin{aligned}
\mathbb{P}(Y=0) &= \mathbb{P}(U=0 \cup V=0) \\
&= \mathbb{P}(U=0) + \mathbb{P}(V=0) - \mathbb{P}(U=0 \cap V=0) \\
&= \mathbb{P}(U=0) + \mathbb{P}(V=0) - \mathbb{P}(U=0)\mathbb{P}(V=0) \\
&= \pi + e^{-\mu} - \pi e^{-\mu} \\
&= \pi + (1-\pi)e^{-\mu}.
\end{aligned} \tag{0.15}$$

In order for Y to take on a value $y > 0$, we must have $U = 1$ and $V = y$. That is,

$$\begin{aligned}
\mathbb{P}(Y=y) &= \mathbb{P}(U=1 \cap V=y) \\
&= \mathbb{P}(U=1)\mathbb{P}(V=y) \\
&= (1-\pi)\frac{e^{-\mu}\mu^y}{y!}, \;\; y = 1, 2, \ldots
\end{aligned} \tag{0.16}$$

Thus, the full probability mass function (pmf) for a ZIP random variable is given by

$$\mathbb{P}(Y=y) = \begin{cases} \pi + (1-\pi)e^{-\mu}, & y = 0 \\ (1-\pi)\frac{e^{-\mu}\mu^y}{y!}, & y = 1, 2, \ldots \end{cases} \tag{0.17}$$

Alternatively, we may write

$$\mathbb{P}(Y=y) = \left(\pi + (1-\pi)e^{-\mu}\right)^{\mathbb{1}_{y=0}} \left((1-\pi)\frac{e^{-\mu}\mu^y}{y!}\right)^{1-\mathbb{1}_{y=0}}, \;\; y = 0, 1, 2, \ldots, \tag{0.18}$$

where $\mathbb{1}_{y=0}$ denotes an indicator function that assumes the value one when $y = 0$ and zero otherwise.

Suppose we observe multiple counts $Y_1, ..., Y_n$ generated independently from Y.[1] The likelihood function for μ and π based on an individual observation $Y_i = y_i$ is simply equal to the pmf for Y evaluated at y_i. That is,

$$L(\mu, \pi; y_i) = \Big(\pi + (1-\pi)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \left((1-\pi)\frac{e^{-\mu}\mu^{y_i}}{y_i!}\right)^{1-\mathbb{1}_{y_i=0}}, \;\; \mu > 0, \;\; \pi \in [0,1]. \tag{0.19}$$

We can safely ignore any multiplicative constant in L without influencing our inferences regarding μ and π. In particular, we can discard the term $\left(\frac{1}{y_i!}\right)^{1-\mathbb{1}_{y_i=0}}$ as it depends only on the observation y_i and is constant with respect to the model parameters. Hence, we may write

$$L(\mu, \pi; y_i) = \Big(\pi + (1-\pi)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \Big((1-\pi)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}, \;\; \mu > 0, \;\; \pi \in [0,1]. \tag{0.20}$$

From this we may derive the corresponding log-likelihood function:

$$\begin{aligned}
\ell(\mu, \pi; y_i) &= \log L(\mu, \pi; y_i) \\
&= \log\left\{\Big(\pi + (1-\pi)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \Big((1-\pi)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}\right\} \\
&= \log\left\{\Big(\pi + (1-\pi)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}}\right\} + \log\left\{\Big((1-\pi)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}\right\} \\
&= \mathbb{1}_{y_i=0}\log\Big(\pi + (1-\pi)e^{-\mu}\Big) + (1-\mathbb{1}_{y_i=0})\log\Big((1-\pi)e^{-\mu}\mu^{y_i}\Big) \\
&= \mathbb{1}_{y_i=0}\log\Big(\pi + (1-\pi)e^{-\mu}\Big) + (1-\mathbb{1}_{y_i=0})\Big(\log(1-\pi) - \mu + y_i\log\mu\Big).
\end{aligned} \tag{0.21}$$

The log-likelihood for the full sample $y_\bullet \equiv (y_1, ..., y_n)$ is then obtained by summing over the

[1] Recall that the parameter of a generic Poisson random variable is defined relative to a fixed length of time during which observations may be recorded and added to the running total. By assuming that all counts come from the same ZIP-distributed random variable, we are implicitly assuming that each count was recorded over the same period of time and thus are on equal footing with one another.

index variable i as follows:

$$
\begin{aligned}
\ell(\mu,\pi;y_\bullet) &= \sum_{i=1}^{n} \ell(\mu,\pi;y_i) \\
&= \sum_{i=1}^{n} \left[\mathbb{1}_{y_i=0} \log\left(\pi + (1-\pi)e^{-\mu}\right) + (1-\mathbb{1}_{y_i=0})\Big(\log(1-\pi) - \mu + y_i \log\mu\Big) \right] \\
&= \sum_{i=1}^{n} \mathbb{1}_{y_i=0} \log\left(\pi + (1-\pi)e^{-\mu}\right) + \sum_{i=1}^{n} (1-\mathbb{1}_{y_i=0})\Big(\log(1-\pi) - \mu + y_i \log\mu\Big) \\
&= \log\left(\pi + (1-\pi)e^{-\mu}\right) \underbrace{\sum_{i=1}^{n} \mathbb{1}_{y_i=0}}_{\text{A}} + \big(\log(1-\pi) - \mu\big) \underbrace{\sum_{i=1}^{n}(1-\mathbb{1}_{y_i=0})}_{\text{B}} + \log\mu \underbrace{\sum_{i=1}^{n}(1-\mathbb{1}_{y_i=0})y_i}_{\text{C}}.
\end{aligned}
\tag{0.22}
$$

The summation in A is counting the number of zero counts in the sample. Let $\bar{\pi}$ represent the proportion of these observed zero counts in the sample so that

$$n\bar{\pi} \equiv \sum_{i=1}^{n} \mathbb{1}_{y_i=0}. \tag{0.23}$$

Similarly, the summation in B is counting the number of nonzero counts in the sample. Since $1-\bar{\pi}$ represents the proportion of the observed nonzero counts, it follows that

$$n(1-\bar{\pi}) \equiv \sum_{i=1}^{n}(1-\mathbb{1}_{y_i=0}). \tag{0.24}$$

Finally, consider the summation in C. Whenever $y_i = 0$, $(1-\mathbb{1}_{y_i=0})y_i = (1-1)\cdot 0 = 0 = y_i$. Whenever $y_i > 0$, $(1-\mathbb{1}_{y_i=0})y_i = (1-0)y_i = y_i$. It follows that $(1-\mathbb{1}_{y_i=0})y_i = y_i$ for all values of y_i. Hence, the summation is simply adding all the observed counts in the sample. Let $\bar{y}$ denote the sample mean so that

$$n\bar{y} \equiv \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(1-\mathbb{1}_{y_i=0})y_i \tag{0.25}$$

Substituting these three expressions on the left in Equation 0.23, Equation 0.24, and Equation 0.25 for their corresponding summations in the final line of Equation 0.22, we arrive at

$$\ell(\mu,\pi;y_\bullet) = n\bar{\pi}\log\left(\pi + (1-\pi)e^{-\mu}\right) + n(1-\bar{\pi})\big(\log(1-\pi) - \mu\big) + n\bar{y}\log\mu. \tag{0.26}$$

Let $\hat{\pi}_\mu$ denote the partial maximum likelihood estimator of π for a given value of μ. Under suitable regularity conditions, easily satisfied in this case since the Poisson distribution belongs to the well-behaved exponential family of distributions, $\hat{\pi}_\mu$ will be the unique value of π (as a function of μ and the data) that solves the critical point equation.

$$\left.\frac{\partial \ell(\mu, \pi)}{\partial \pi}\right|_{\pi=\hat{\pi}_\mu} \equiv 0. \tag{0.27}$$

To find it, we start by differentiating both sides of Equation 0.26:

$$\frac{\partial \ell(\mu, \pi)}{\partial \pi} = \frac{n\bar{\pi}}{\pi + (1-\pi)e^{-\mu}}(1-e^{-\mu}) - \frac{n(1-\bar{\pi})}{1-\pi} = n\left[\frac{\bar{\pi}(1-e^{-\mu})}{\pi + (1-\pi)e^{-\mu}} - \frac{1-\bar{\pi}}{1-\pi}\right].$$

Evaluating at $\pi = \hat{\pi}_\mu$ and plugging the result into Equation 0.26, we have

$$n\left[\frac{\bar{\pi}(1-e^{-\mu})}{\hat{\pi}_\mu + (1-\hat{\pi}_\mu)e^{-\mu}} - \frac{1-\bar{\pi}}{1-\hat{\pi}_\mu}\right] = 0. \tag{0.28}$$

When we solve for $\hat{\pi}_\mu$ (see the appendix for the full derivation), we get

$$\hat{\pi}_\mu = \frac{\bar{\pi} - e^{-\mu}}{1 - e^{-\mu}}. \tag{0.29}$$

This formula for $\hat{\pi}_\mu$ leads to a negative estimate of π when $\bar{\pi} < e^{-\mu}$, which is clearly undesirable since π is meant to be interpreted as a probability. To correct it, we can modify $\hat{\pi}_\mu$ slightly by setting it equal to the value in the interval $[0, 1]$ that maximizes ℓ whenever $\bar{\pi} < e^{-\mu}$. Since π represents the probability of excess zero counts in the sample, and this situation can only occur when the observed proportion of zero counts (i.e. $\bar{\pi}$) in the sample is less than what we would expect if our Poisson variable had no zero-inflation at all[2], it is unsurprising that this value turns out simply to be zero (see the appendix for a formal proof). Hence, our full formula for the partial-MLE of π given a particular value of μ is as follows:

$$\hat{\pi}_\mu = \begin{cases} \frac{\bar{\pi} - e^{-\mu}}{1 - e^{-\mu}}, & \bar{\pi} \geq e^{-\mu} \\ 0, & \bar{\pi} < e^{-\mu} \end{cases}$$

As we hinted at above, the condition that $\bar{\pi} \geq e^{-\mu}$ could be considered a bellwether for zero-inflation, as it indicates an empirical excess of zero counts relative to what we would expect for a non-zero-inflated random variable. We refer to the condition that $\bar{\pi} \geq e^{-\mu}$ as the *zero-inflated condition*.

Importance Sampling

Let $p(\theta)$ denote a prior distribution for a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ and $L(\theta; X)$ the likelihood function of our model based on data X. The posterior distribution for θ is given by $\pi(\theta|X) =$

[2]Recall that $e^{-\mu}$ is the probability of observing a zero under a non-zero-inflated Poisson random variable with rate parameter μ.

$cL(\theta;X)p(\theta)$, where $c = \left(\int_\Theta L(\theta;X)p(\theta)d\theta\right)^{-1} < \infty$. Suppose we have another function $f(\theta) > 0$ for all $\theta \in \Theta$ and we are interested in estimating the expectation of this function with respect to the distribution of p. Call this value μ. Then we have

$$\begin{aligned}
\mu &= \mathrm{E}_p(f(\theta)) \\
&= \int_\Theta f(\theta)p(\theta)d\theta \\
&= \int_\Theta \frac{f(\theta)}{cL(\theta;X)} cL(\theta;X)p(\theta)d\theta \\
&= \int_\Theta \frac{f(\theta)}{cL(\theta;X)} \pi(\theta|X)d\theta \\
&= \mathrm{E}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right).
\end{aligned}$$

The *importance sampling estimator* for μ is

$$\hat{\mu}_\pi = \frac{1}{R}\sum_{i=1}^{R} \frac{f(\theta_i)}{cL(\theta_i;X)}, \quad \theta_i \sim \pi.$$

Note that $\hat{\mu}_\pi$ is unbiased, i.e.

$$\begin{aligned}
\mathrm{E}_\pi(\hat{\mu}_\pi) &= \mathrm{E}_\pi\left(\frac{1}{R}\sum_{i=1}^{R} \frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R}\sum_{i=1}^{R} \mathrm{E}_\pi\left(\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R}\sum_{i=1}^{R} \mu \\
&= \frac{1}{R}R\mu \\
&= \mu,
\end{aligned}$$

and by the law of large numbers converges in distribution to μ, i.e.

$$\hat{\mu}_\pi \to \mu \;\text{ as }\; R \to \infty.$$

The variance of $\widehat{\mu}_\pi$ is given by

$$
\begin{aligned}
\mathrm{Var}_\pi(\widehat{\mu}_\pi) &= \mathrm{Var}_\pi\left(\frac{1}{R}\sum_{i=1}^{R}\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R^2}\sum_{i=1}^{R}\mathrm{Var}_\pi\left(\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R^2}\sum_{i=1}^{R}\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R^2}R\cdot\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R}\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R}\left\{\mathrm{E}_\pi\left[\left(\frac{f(\theta)}{cL(\theta;X)}\right)^2\right] - \left[\mathrm{E}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right)\right]^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\left(\frac{f(\theta)}{cL(\theta;X)}\right)^2\pi(\theta|X)d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2}{c^2L(\theta;X)^2}cL(\theta;X)p(\theta)d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2p(\theta)}{cL(\theta;X)}d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2p(\theta)^2}{cL(\theta;X)p(\theta)}d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)}d\theta - \mu^2\right\} \\
&= \frac{\sigma_\pi^2}{R},
\end{aligned}
$$

where

$$
\sigma_\pi^2 = \int_\Theta\frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)}d\theta - \mu^2.
$$

Some clever rearranging and substituting allows us to rewrite it as

$$\begin{aligned}
\sigma_\pi^2 &= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)} d\theta - \mu^2 \\
&= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)} d\theta - 2\mu^2 + \mu^2 \\
&= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)} d\theta - 2\mu \int_\Theta f(\theta)p(\theta)d\theta + \mu^2 \int_\Theta \pi(\theta|X)d\theta \\
&= \int_\Theta \left(\frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)} - 2\mu f(\theta)p(\theta) + \mu^2\pi(\theta|X) \right) d\theta \\
&= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2 - 2\mu f(\theta)p(\theta)\pi(\theta|X) + \mu^2\pi(\theta|X)^2}{\pi(\theta|X)} d\theta \\
&= \int_\Theta \frac{\big(f(\theta)p(\theta) - \mu\pi(\theta|X)\big)^2}{\pi(\theta|X)} d\theta.
\end{aligned}$$

We can also write

$$\begin{aligned}
\sigma_\pi^2 &= \int_\Theta \frac{\big(f(\theta)p(\theta) - \mu\pi(\theta|X)\big)^2}{\pi(\theta|X)} d\theta \\
&= \int_\Theta \left(\frac{f(\theta)p(\theta) - \mu\pi(\theta|X)}{\pi(\theta|X)} \right)^2 \pi(\theta|X) d\theta \\
&= \mathrm{E}_\pi \left[\left(\frac{f(\theta)p(\theta) - \mu\pi(\theta|X)}{\pi(\theta|X)} \right)^2 \right].
\end{aligned}$$

Because the θ_i are sampled from π, the natural variance estimate is

$$\hat{\sigma}_\pi^2 = \frac{1}{R} \sum_{i=1}^{R} \left(\frac{f(\theta_i)}{cL(\theta_i;X)} - \hat{\mu}_\pi \right)^2 = \frac{1}{R} \sum_{i=1}^{R} (w_i f(\theta_i) - \hat{\mu}_\pi)^2,$$

where $w_i = \frac{1}{cL(\theta_i;X)}$.

$$
\begin{aligned}
\sigma_\pi^2 + \mu &= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2}{\pi(\theta|X)} d\theta \\
&= \int_\Theta \frac{\big(f(\theta)p(\theta)\big)^2}{cL(\theta;X)p(\theta)} d\theta \\
&= \int_\Theta \frac{f(\theta)^2}{cL(\theta;X)} p(\theta) d\theta \\
&= \mathrm{E}_p\left(\frac{f(\theta)^2}{cL(\theta;X)}\right) \\
&= \mathrm{E}_\pi\left(\frac{f(\theta)^2}{c^2L(\theta;X)^2}\right).
\end{aligned}
$$

Self-normalized importance sampling

$\pi(\theta|X) = cL(\theta;X)p(\theta)$, $c > 0$ unknown.
$p_u(\theta) = ap(\theta)$, $a > 0$ unknown.
$p_u(\theta) = bp(\theta)$, $a > 0$ unknown.

$$
\tilde{\mu}_\pi =
$$

Appendix

Claim: $\pi + (1-\pi)e^{-\mu} \geq e^{-\mu}$ for all $\pi \in [0,1]$.
Proof:
$$
$$

$$
\begin{aligned}
& n\left[\frac{\bar{\pi}(1-e^{-\mu})}{\hat{\pi}_\mu+(1-\hat{\pi}_\mu)e^{-\mu}}-\frac{1-\bar{\pi}}{1-\hat{\pi}_\mu}\right]=0 \\
& \implies \frac{\bar{\pi}(1-e^{-\mu})}{\hat{\pi}_\mu+(1-\hat{\pi}_\mu)e^{-\mu}}=\frac{1-\bar{\pi}}{1-\hat{\pi}_\mu} \\
& \implies \frac{(1-\hat{\pi}_\mu)(1-e^{-\mu})}{\hat{\pi}_\mu+(1-\hat{\pi}_\mu)e^{-\mu}}=\frac{1-\bar{\pi}}{\bar{\pi}} \\
& \implies \frac{\hat{\pi}_\mu+(1-\hat{\pi}_\mu)e^{-\mu}}{(1-\hat{\pi}_\mu)(1-e^{-\mu})}=\frac{\bar{\pi}}{1-\bar{\pi}} \\
& \implies \frac{\hat{\pi}_\mu}{(1-\hat{\pi}_\mu)(1-e^{-\mu})}+\frac{(1-\hat{\pi}_\mu)e^{-\mu}}{(1-\hat{\pi}_\mu)(1-e^{-\mu})}=\frac{\bar{\pi}}{1-\bar{\pi}} \\
& \implies \frac{\hat{\pi}_\mu}{(1-\hat{\pi}_\mu)(1-e^{-\mu})}=\frac{\bar{\pi}}{1-\bar{\pi}}-\frac{e^{-\mu}}{1-e^{-\mu}} \\
& \implies \frac{\hat{\pi}_\mu}{1-\hat{\pi}_\mu}=\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu} \\
& \implies \frac{\hat{\pi}_\mu}{1-\hat{\pi}_\mu}=\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu} \\
& \implies \hat{\pi}_\mu=\frac{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu}}{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu}+1} \\
& \implies \hat{\pi}_\mu=\frac{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu}}{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})+(1-e^{-\mu})} \\
& \implies \hat{\pi}_\mu=\frac{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu}}{(\frac{\bar{\pi}}{1-\bar{\pi}}+1)(1-e^{-\mu})} \\
& \implies \hat{\pi}_\mu=\frac{\frac{\bar{\pi}}{1-\bar{\pi}}(1-e^{-\mu})-e^{-\mu}}{\frac{1}{1-\bar{\pi}}(1-e^{-\mu})} \\
& \implies \hat{\pi}_\mu=\frac{\bar{\pi}(1-e^{-\mu})-(1-\bar{\pi})e^{-\mu}}{1-e^{-\mu}} \\
& \implies \hat{\pi}_\mu=\frac{\bar{\pi}-e^{-\mu}}{1-e^{-\mu}}
\end{aligned}
$$