Title
A Bayesian Large Deviations Probabilistic Interpretation and Justification of Empirical Likelihood

Permalink
https://escholarship.org/uc/item/1z012014

Authors
Grendar, Marian
Judge, George G.

Publication Date
2007-04-23
A BAYESIAN LARGE DEVIATIONS PROBABILISTIC INTERPRETATION AND JUSTIFICATION OF EMPIRICAL LIKELIHOOD

By Marian Grendár* and George Judge

Bel University and University of California, Berkeley

In this paper we demonstrate, in a parametric Estimating Equations setting, that the Empirical Likelihood (EL) method is an asymptotic instance of the Bayesian non-parametric Maximum-A-Posteriori approach. The resulting probabilistic interpretation and justification of EL rests on Bayesian non-parametric consistency in $L$-divergence.

1. Introduction. Empirical Likelihood (EL) is first and foremost a method of inference; cf. [17], but the EL criterion has also been used as a regularization criterion in an Estimating Equations (EE) context; cf. [19]. We provide a Bayesian probabilistic justification and interpretation of EL in the context of EE. We show that EL is an asymptotic instance of the non-parametric Maximum-A-Posteriori (MAP) approach, when embedded in EE. The probabilistic argument also implies that, for finite samples, one can use the MAP-EE estimator.

Despite several previous attempts, EL lacks a probabilistic interpretation. An early attempt to replace likelihood by EL in the parametric Bayesian context is [15]. Owen ([17], Ch. 9) notes a similarity between EL and Bayesian bootstrap [21]. Bayesian bootstrap is a bootstrap performed in non-parametric Bayesian context, where a Dirichlet prior is assumed over a set of sampling distributions. In [3], this type of Bayesian bootstrap was considered in an EE context. Schennach [22] uses a specific prior over a set of sampling distributions to get a Bayesian procedure that admits an operational form similar to EL. In [20] a different prior over the set of probability measures (i.e., a non-parametric prior) is considered, and in this way a group of EL-like methods is obtained. We highlight two features that distinguish our argument from that of Schennach [22] and Ragusa [20] as well as from the Bayesian bootstrap: 1) though Schennach as well as Ragusa consider the

*Supported by VEGA 1/3016/06 grant.
AMS 2000 subject classifications: Primary 62G05, 62C10; secondary 60F10
Keywords and phrases: Maximum Non-parametric Likelihood, Estimating Equations, Bayesian non-parametric consistency, Sanov Theorem for Sampling Distributions, $L$-divergence

1For exposition, see [16], Chap. 12 and 13 and [17], Sect. 3.5.
Bayesian non-parametric setting, they study a particular prior distribution. Consequently, the authors obtain in a non-parametric Bayesian way methods that are similar to EL. One could in principle imagine other ways of constructing a prior over a set of probability measures, that could admit an operational representation similar to EL; 2) These authors do not perform asymptotic considerations of Bayesian non-parametric consistency with their procedures. We show, loosely put, that regardless of the prior used over a set of data sampling distributions, asymptotically the Bayesian non-parametric Maximum A-Posteriori selection turns into EL.

The probabilistic interpretation of EL we propose, is based on Sanov’s Theorem for Sampling Distributions (cf. [1], [2], [6], [8], [9] for independent developments in this direction). Sanov’s Theorem for Sampling Distributions (abbreviated LST due to the key role that $L$-divergence plays there) is the basic result of Large Deviations for sampling distributions. We stress that LST should not be confused with the standard Sanov’s Theorem for Empirical Measures [5]. The starting point of our quest for a probabilistic justification of EL was a key observation due to Kitamura and Stutzer [12]. They note that the Sanov’s Theorem for Empirical Measures and its Corollary the Conditional Limit Theorem provide a probabilistic justification of the Relative Entropy Maximization method [4] and can also be applied in an EE context, to provide a probabilistic underpinning of Maximum Entropy Empirical Likelihood (MEEL) method. However, Sanov’s Theorem for Empirical Measures cannot be used to justify the Empirical Likelihood method. In this paper we show how Sanov’s Theorem for Sampling Distributions provides both a probabilistic interpretation and a justification of EL, as an asymptotic instance of the Bayesian MAP, when embedded into EE context.

While we phrase the interpretation and justification of EL in an EE context, it holds wherever the underlying Bayesian non-parametric consistency in $L$-divergence holds.

### 2. Bayesian Probabilistic Interpretation and Justification of EL.

Let $X$ be a discrete random variable with probability mass function (pmf) $r(X; \theta)$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $X$ take on $m$ values from $\mathcal{X}$.

Assume that a researcher is not willing to specify the data sampling distribution $q(X; \theta)$, but is only willing to specify some of the underlying features of the data-sampling distribution. These features can be characterized by $J$ unbiased estimating equations $E_X h_j(X; \theta) = 0$, $j = 1, 2, \ldots, J$, where $J$ might be different than the number of parameters $k$ that parameterize the data sampling distribution. Moreover, in general, the data sampling distri-
Bayesian Probabilistic Justification of EL

...distribution \( r(X; \theta) \) need not belong to the model set \( \Phi(\Theta) \equiv \bigcup_{\theta \in \Theta} \Phi(\theta) \), where \( \Phi(\theta) \equiv \{q(\theta) : \sum_{i=1}^{m} q_i(\theta) h_j(x_i; \theta) = 0, j = 1, 2, \ldots, J\} \).

Assume for the sake of simplicity that \( \Theta = \{\theta_1, \theta_2\} \). Suppose a non-parametric Bayesian using these EE puts a prior \( \pi(q; \theta) \) over the set of data-sampling distributions \( \Phi(\Theta) \) and arrives at the posterior \( \pi(q | x^n; \theta) \) after a random sample \( x^n = x_1, x_2, \ldots, x_n \) is drawn from \( r(X; \theta_0) \). Then, a Bayesian may use the Maximum-A-Posteriori sampling distribution

\[
\hat{q}(X; \hat{\theta}) \equiv \arg \sup_{\theta \in \Theta} \sup_{q \in \Phi(\theta)} \pi(q | x^n; \theta)
\]

to estimate the true \( r(X; \theta_0) \). The \( \hat{\theta} \) of \{\theta_1, \theta_2\} for which MAP is attained is taken as the estimator of \( \theta \). We thus have a hybrid estimator \( \hat{q}(X; \hat{\theta}) \) given by

\[
(1a) \quad \hat{\theta} = \arg \sup_{\theta \in \Theta} \pi(\hat{q}(X; \theta) | x^n; \theta),
\]

where

\[
(1b) \quad \hat{q}(X; \theta) = \arg \sup_{q \in \Phi(\theta)} \pi(q | x^n; \theta).
\]

We call the parametric component \( \hat{\theta} \) of the hybrid estimator, the MAP-EE estimator.

There is a strong probabilistic justification for picking a \( \hat{q}(\hat{\theta}) \), that rests on \( L \)-divergence consistency of non-parametric Bayesian procedures. The asymptotic consistency in \( L \)-divergence is a direct consequence of Sanov’s Theorem for Sampling Distributions; cf. Sect. 3. Loosely put, \( L \)-divergence consistency means that, as \( n \to \infty \), the posterior probability \( \pi(q \in \Phi | x^n) \) concentrates almost surely on the MAP sampling distribution. This distribution is asymptotically equivalent to the \( L \)-projection \( \hat{q} \) of \( r \) on \( \Phi \)

\[
\hat{q} \equiv \arg \inf_{q \in \Phi} - \sum_{i=1}^{m} r_i \log q_i.
\]

In other words, the consistency result demonstrates that data sampling distributions with Maximum-A-Posteriori probability, are asymptotically a posteriori the only ones possible. Hence, in general, in a non-parametric Bayesian setting, the consistency requirement precludes selecting say a Mean-A-Posteriori data sampling distribution (or a distribution that minimizes a discrepancy measure non-trivially different than \( L \)-divergence), since, it

\[\text{2There arg stands for 'argument', i.e., 'point of'.}\]
would be a selection of an asymptotically a posteriori zero-probable data sampling distribution.

Observe that, since the $L$-projection is just the asymptotic form of Maximum Non-parametric Likelihood, the prior does not matter as $n \to \infty$. Hence, by $L$-divergence consistency as $n \to \infty$, (1) becomes:

\[
\hat{\theta} = \arg \sup_{\theta \in \Theta} \sum_{i=1}^{m} r_i \log \hat{q}(X_i; \theta),
\]

where

\[
\hat{q}(\theta) = \arg \sup_{q \in \Phi(\theta)} \sum_{i=1}^{m} r_i \log q(X_i; \theta).
\]

We call this estimator the Maximum Non-Parametric Likelihood with parametric Estimating Equations (MNPL-EE).

One can view MNPL-EE as an asymptotic instance of MAP-EE or as MAP-EE with an uninformative prior over the entire $\Phi(\theta)$. However, it is equally legitimate to view MNPL-EE as a self-standing method, which implies that, for finite $n$, one selects

\[
\hat{\theta} = \arg \sup_{\theta \in \Theta} \sum_{i=1}^{m} \nu_i^n \log \hat{q}(X_i; \theta),
\]

where

\[
\hat{q}(\theta) = \arg \sup_{q \in \Phi(\theta)} \sum_{i=1}^{m} \nu_i^n \log q(X_i; \theta),
\]

and $\nu^n$ is the empirical pmf induced by the random sample $x^n$. All of these views comply with the Bayesian non-parametric consistency of $L$-divergence.

Note that, in the case of a discrete random variables, (3) can also be written in the unconcentrated form:

\[
\hat{\theta} = \arg \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log \hat{q}(X_i; \theta),
\]

where

\[
\hat{q}(\theta) = \arg \sup_{q \in \Phi(\theta)} \frac{1}{n} \sum_{i=1}^{n} \log q(X_i; \theta).
\]

The estimator (3'), which we call the n-MNPL-EE estimator, is thus a discrete-case instance of the Empirical Likelihood estimator.
Turning now to the case of continuous $X$, observe that Bayesian non-parametric consistency in $L$-divergence still holds, leading to the MNPL-EE estimator $\hat{\theta}$ of $\theta$ as $n \to \infty$:

\[
\hat{\theta} = \arg \sup_{\theta \in \Theta} \int r(X; \theta_0) \log \hat{q}(X; \theta),
\]

where

\[
\hat{q}(X; \theta) = \arg \sup_{q \in \Phi(\theta)} \int r(X; \theta_0) \log q(X; \theta).
\]

If, in the continuous case, one views the MNPL-EE estimation method as a self-standing method rather than an asymptotic instance of the continuous-case MAP-EE, then one faces the problem of finding a finite-$n$ counterpart of (4). Calculating (3) is clearly not viable, but evaluation of (3') is feasible with the use of a technique suggested by Owen ([17], [16], Ch. 12.1.2.a): let the sample $x_1, x_2, \ldots, x_n$, become a support of an auxiliary random variable $S$. The finite-$n$ form of (4) is thus

\[
\hat{\theta} = \arg \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log \hat{q}(x_i; \theta),
\]

where

\[
\hat{q}(X; \theta) = \arg \sup_{q \in \Phi(\theta)} \frac{1}{n} \sum_{i=1}^{n} \log q(x_i; \theta),
\]

and $\Phi(\theta) = \{ q : \sum_{i=1}^{n} q(x_i; \theta) b_j(x_i; \theta) = 0, j = 1, 2, \ldots, J \}$. The resulting estimator is identical to the EL estimator. One can also arrive at the continuous-case EL estimator (5) without using Owen’s technique; cf. [12].

In the next section, a formal statement of the Large Deviations result is given.

3. Bayesian non-parametric consistency in $L$-divergence. Let $\mathcal{M} = \{ q_1, q_2, \ldots \}$ be a countable set of probability density functions with respect to the Lebesgue measure. Suppose a Bayesian puts a strictly positive prior probability mass function $\pi(\cdot)$ on this set. Let $r$ be the true sampling distribution of a random sample $X^n \triangleq X_1, X_2, \ldots, X_n$. Provided that $r \in \mathcal{M}$, the posterior distribution $\pi(\cdot|X^n = x^n)$ over $\mathcal{M}$ is expected to concentrate in a neighborhood of the true sampling distribution $r$ as $n \to \infty$. The conditions under which this indeed happens is a subject of Bayesian non-parametric consistency investigations. Surveys regarding this issue include [7], [23], [24].
Let $\mathcal{M}^e \triangleq \{q : q \in \mathcal{M}, \pi(q) > 0\}$ be the support of a prior pmf which does not necessarily contain $r$. Thus, we are also interested in Bayesian consistency under misspecification, i.e., when $\pi(r) = 0$.

For two densities $p, q$ with respect to the Lebesgue measure $\lambda$, the $I$-divergence $I(p||q) \triangleq \int p \log(p/q)$. The $L$-divergence $L(q||p)$ of $q$ with respect to $p$ is defined as $L(q||p) \triangleq -\int p \log q$. The $L$-projection $\hat{q}$ of $p$ on $\mathcal{Q}$ is $\hat{q} \triangleq \arg \inf_{q \in \mathcal{Q}} L(q||p)$, where $\mathcal{Q}$ is a set of probability densities defined on the same support. The value of $L$-divergence at an $L$-projection of $p$ on $\mathcal{Q}$ is denoted by $L(\mathcal{Q}||p)$. It is implicitly assumed that all the relevant values of $L$-divergence at $L$-projections are finite. Note that $L$-projection is formally identical with the reverse $I$-projection; cf. [8] for a discussion of the pros and cons of stating Sanov’s Theorem for Sampling Distributions in terms of $L$-projection rather than reverse $I$-projection.

Bayesian non-parametric consistency in $L$-divergence is a direct consequence (corollary) of Sanov’s Theorem for Sampling Distributions (LST).

**LST [9]** Let $\mathcal{N} \subset \mathcal{M}^e$. As $n \to \infty$,

$$\Pr \left( \frac{1}{n} \log \pi(q \in \mathcal{N}|x^n) \to -\{L(\mathcal{M}^e||r) - L(\mathcal{N}||r)\} \right) = 1.$$ 

**Proof** Let $l_n(q) \triangleq \exp(\sum_{i=1}^{n} \log q(X_i))$, $l_n(A) \triangleq \sum_{q \in A} l_n(q)$, and $\rho_n(q) \triangleq \pi(q)l_n(q)$, $\rho_n(A) \triangleq \sum_{q \in A} \rho_n(q)$. In this notation $\pi(q \in \mathcal{N}|x^n) = \frac{\rho_n(\mathcal{N})}{\rho_n(\mathcal{M}^e)}$. The posterior probability is bounded above and below as follows:

$$\frac{\hat{\rho}_n(\mathcal{N})}{\hat{l}_n(\mathcal{M}^e)} \leq \pi(q \in \mathcal{N}|x^n) \leq \frac{\hat{l}_n(\mathcal{N})}{\hat{\rho}_n(\mathcal{M}^e)},$$

where $\hat{l}_n(A) \triangleq \sup_{q \in A} l_n(q)$, $\hat{\rho}_n(A) \triangleq \sup_{q \in A} \rho_n(q)$.

$$\frac{1}{n}(\log \hat{l}_n(\mathcal{N}) - \log \hat{\rho}_n(\mathcal{M}^e))$$

converges with probability one to $L(\mathcal{N}||r) - L(\mathcal{M}^e||r)$. This is the same as the 'point' of a.s. convergence of $\frac{1}{n} \log$ of the lower bound. \qed

Let $\mathcal{N}^C(\mathcal{M}^e) \triangleq \{q : L(\mathcal{M}^e||r) - L(q||r) > \epsilon, q \in \mathcal{M}^e\}$ for $\epsilon > 0$. Then, we have the consistency result:

**Corollary** Let there be a finite number of $L$-projections of $r$ on $\mathcal{M}^e$. For any $\epsilon > 0$, as $n \to \infty$,

$$\Pr \left( \pi(q \in \mathcal{N}^C(\mathcal{M}^e)|x^n) \to 0 \right) = 1.$$
$L$-divergence consistency shows that, as $n \to \infty$, the posterior measure 
$\pi(q \in N \mid x^n)$ concentrates in an arbitrarily small neighborhood of the $L$
-projection(s) of $r$ on $N$.

To the best of our knowledge, Bayesian Sanov’s Theorem has not been
developed for a general setting. In [1], [2], the Theorem is proved for a con-
tinuous set of sampling distributions with a finite support $\mathcal{X}$ and in [6] under
the additional assumption that $\pi(r) > 0$. Without using the Large Devia-
tions approach, Kleijn and van der Vaart [14] developed sufficient conditions
for posterior concentration on $L$-projection(s) for the case of continuous prior
over a set of continuous sampling distributions. In the same general setting,
Bayesian Sanov’s Theorem for $n$-data sampling distributions is studied in
[8].

A simple illustrative example is presented in the Appendix.

4. Summary. Despite several attempts, the Empirical Likelihood (EL)
criterion lacks a probabilistic underpinning. In this paper we have, within
an Estimating Equations (EE) context, demonstrated that Empirical Like-
lihood is an asymptotic instance of the non-parametric Bayesian Maximum-
A-Posteriori approach. This interpretation arises directly from $L$-divergence
consistency of Bayesian non-parametric methods and provides a probabilis-
tic interpretation and justification of the EL criterion function. Indeed, in
this context, application of any other criterion (e.g. Euclidean likelihood)
for construction of empirical estimator, would, in general, violate Bayesian

The Large Deviations approach to Bayesian non-parametric consistency
(i.e., Bayesian Sanov’s Theorem) is capable of picking up pseudo-metrics
under which it is natural to study consistency. For $iid$-sampling, the natural
discrepancy measure turns out to be $L$-divergence. In general, for a sampling
scheme other than $iid$, a discrepancy measure other than $L$-divergence, will
govern the posterior concentration. In this sense, EL is limited to the $iid$
sampling scenario. Study of non-$iid$ settings is subject of current research.

Appendix. This example is not meant to illustrate convergence, but to
accompany the discussion of Section 2.

Let $X$ be a discrete random variable with support $\mathcal{X} = \{1, 2, 3, 4\}$ and a
pmf $r(X; \theta)$ that is unknown to us. Let $\Phi(\theta) = \{q(\theta) : \sum_\mathcal{X} q_i(\theta)(x_i - \theta) = 0\}$
and let $\Theta = \{1.5, 3.2\}$. To make $\Phi(\theta)$ conformant with LST (i.e., countable),
we arbitrarily restrict the set $\Phi(\theta)$ to those pmf’s that are rational and have
Table 1

Data related to the Example.

<table>
<thead>
<tr>
<th></th>
<th>(10q_1)</th>
<th>(10q_2)</th>
<th>(10q_3)</th>
<th>(10q_4)</th>
<th>(\pi(q))</th>
<th>(\pi(q \mid \nu^{20}))</th>
<th>(\sum \nu^{20}\log q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\theta_1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.3014</td>
<td>0.3365 (\ast)</td>
<td>-1.8867</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0.1507</td>
<td>0.3061</td>
<td>-1.2883 (\ast)</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.1005</td>
<td>0.1680</td>
<td>-1.4829</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.0753</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.0603</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>(\theta_2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0.0502</td>
<td>0.0999</td>
<td>-1.310</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0.0431</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>0.0377</td>
<td>0.0506</td>
<td>-1.7016</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0.0335</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0.0301</td>
<td>0.0390</td>
<td>-1.7393</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0.0274</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0.0251</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0.0232</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>0.0215</td>
<td>0</td>
<td>-\infty</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>0.0201</td>
<td>0</td>
<td>-\infty</td>
</tr>
</tbody>
</table>

all denominators equal to 10. Then, there are 5 such pmf’s for \(\theta_1 = 1.5\) and 10 such pmf’s for \(\theta_2 = 3.2\) (see Table 1). Thus, there are 15 pmf’s in \(\Phi(\Theta)\), which we index with \(i\). Suppose a non-parametric Bayesian puts the prior \(\pi(q; \theta) \propto \frac{1}{q}\) over the data-sampling distributions from Table 1. Then, if a random sample of size \(n = 20\) arrives and induces type \(\nu^{20} = [4, 9, 7, 0]/20\), the Bayesian finds posterior probabilities \(\pi(q \mid \nu^{20}, \theta)\); see Table 1. The MAP sampling distribution is \(\hat{q}(\theta_1) = [8, 0, 1, 1]/10\), belonging to those sources that satisfy the EE for \(\theta_1 = 1.5\). Hence, the MAP-EE estimator of the parameter \(\theta\) is \(\hat{\theta} = 1.5\). Values of the likelihood \(\sum \nu^{20}\log q\) are also given in Table 1. The n-MNPL-EE estimator is the second sampling distribution \(\hat{q} = [7, 2, 0, 1]/10\), which belongs to the \(\theta_1\) group. Hence, the n-MNPL-EE estimator of \(\theta\) is \(\hat{\theta} = 1.5\).

Acknowledgement. We are grateful to Nicole Lazar who gave us valuable feedback that encouraged us to write this note and to make our argument more transparent. We also want to thank Douglas Miller, Art Owen, Dale Poirier, Jing Qin and Giuseppe Ragusa for valuable comments and suggestions.
References.
