UC Berkeley
CUDARE Working Papers

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[1]. 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

[1]For 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 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 *L*ST due to the key role that *L*-divergence plays there) is the basic result of Large Deviations for sampling distributions. We stress that *L*ST 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-

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

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[2]

$$\hat{q}(X;\hat{\theta}) \triangleq \arg\sup_{\theta\in\Theta} \sup_{q\in\Phi(\theta)} \pi(q\,|\,x^n;\theta)$$

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 θ. We thus have a hybrid estimator $\hat{q}(X;\hat{\theta})$ given by

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

where

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

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 Φ

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

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

---

[2]There 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:

$$(2a) \qquad \hat{\theta} = \arg\sup_{\theta \in \Theta} \sum_{i=1}^{m} r_i \log \hat{q}(X_i; \theta),$$

where

$$(2b) \qquad \hat{q}(\theta) = \arg\sup_{q \in \Phi(\theta)} \sum_{i=1}^{m} r_i \log q(X_i; \theta).$$

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

$$(3a) \qquad \hat{\theta} = \arg\sup_{\theta \in \Theta} \sum_{i=1}^{m} \nu_i^n \log \hat{q}(X_i; \theta),$$

where

$$(3b) \qquad \hat{q}(\theta) = \arg\sup_{q \in \Phi(\theta)} \sum_{i=1}^{m} \nu_i^n \log q(X_i; \theta),$$

and ν^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:

$$(3a') \qquad \hat{\theta} = \arg\sup_{\theta \in \Theta} \frac{1}{n} \sum_{l=1}^{n} \log \hat{q}(X_l; \theta),$$

where

$$(3b') \qquad \hat{q}(\theta) = \arg\sup_{q \in \Phi(\theta)} \frac{1}{n} \sum_{l=1}^{n} \log q(X_l; \theta).$$

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 θ as $n \to \infty$:

(4a)
$$\hat{\theta} = \arg \sup_{\theta \in \Theta} \int r(X; \theta_0) \log \hat{q}(X; \theta),$$

where

(4b)
$$\hat{q}(X; \theta) = \arg \sup_{q \in \Phi(\theta)} \int r(X; \theta_0) \log q(X; \theta).$$

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

(5a)
$$\hat{\theta} = \arg \sup_{\theta \in \Theta} \frac{1}{n} \sum_{l=1}^{n} \log \hat{q}(x_l; \theta),$$

where

(5b)
$$\hat{q}(X; \theta) = \arg \sup_{q \in \Phi(\theta)} \frac{1}{n} \sum_{l=1}^{n} \log q(x_l; \theta),$$

and $\Phi(\theta) = \{q : \sum_{l=1}^{n} q(x_l; \theta) h_j(x_l; \theta) = 0, j = 1, 2, \ldots, J\}$. 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 λ, 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 $l_n(q) \triangleq \exp(\sum_{l=1}^n \log q(X_l))$, $l_n(A) \triangleq \sum_{q \in A} l_n(q)$, 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. $\square$

Let $\mathcal{N}_\epsilon^C(\mathcal{M}^e) \triangleq \{q : L(\mathcal{M}^e||r) - L(q||r) > \epsilon, q \in \mathcal{M}^e\}$ 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 $n \to \infty$,*

$$\Pr \left(\pi(q \in \mathcal{N}_\epsilon^C(\mathcal{M}^e)|x^n) \to 0 \right) = 1.$$

L-divergence consistency shows that, as $n \to \infty$, the posterior measure $\pi(q \in \mathcal{N} \,|\, x^n)$ concentrates in an arbitrarily small neighborhood of the L-projection(s) of r on $\mathcal{N}$.

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 continuous 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 Deviations 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 Likelihood 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 probabilistic 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 asymptotic consistency; cf. [10], [11] for further discussion.

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.

	i	$10q_1$	$10q_2$	$10q_3$	$10q_4$	$\pi(q)$	$\pi(q \mid \nu^{20})$	$\sum \nu^{20} \log q$
	1	8	0	1	1	0.3014	0.3365 ⋆	-1.8867
	2	7	2	0	1	0.1507	0.3061	-1.2883 ⋆
θ_1	3	7	1	2	0	0.1005	0.1680	-1.4829
	4	6	3	1	0	0.0753	0	$-\infty$
	5	5	5	0	0	0.0603	0	$-\infty$
	6	2	1	0	7	0.0502	0.0999	-1.310
	7	2	0	2	6	0.0431	0	$-\infty$
	8	1	2	1	6	0.0377	0.0506	-1.7016
	9	0	4	0	6	0.0335	0	$-\infty$
	10	1	1	3	5	0.0301	0.0390	-1.7393
θ_2	11	0	3	2	5	0.0274	0	$-\infty$
	12	1	0	5	4	0.0251	0	$-\infty$
	13	0	2	4	4	0.0232	0	$-\infty$
	14	0	1	6	3	0.0215	0	$-\infty$
	15	0	0	8	2	0.0201	0	$-\infty$

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 $\pi(q; \theta) \propto \frac{1}{i}$ 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 θ 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 θ_1 group. Hence, the n-MNPL-EE estimator of θ is $\hat{\theta} = 1.5$.

References.

[1] Ben-Tal, A., Brown, D. E. and R. L. Smith. (1987). Posterior convergence under incomplete information. Tech. rep. 87-23. U. of Michigan, Ann Arbor.

[2] Ben-Tal, A., Brown, D. E. and R. L. Smith. (1988). Relative Entropy and the convergence of the posterior and empirical distributions under incomplete and conflicting information. Tech. rep. 88-12. U. of Michigan.

[3] Chamberlain, G. and G. W. Imbens (2003). Nonparametric applications of bayesian inference. *Jour. Business Econ. Stat.* 21:12-18.

[4] Csiszár I. (1996). MaxEnt, mathematics and information theory, in *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver (eds.), pp. 35-50, Kluwer, 1996.

[5] Dembo, A. and O. Zeitouni. (1998). *Large Deviations Techniques and Applications.* New York:Springer-Verlag.

[6] Ganesh, A. and N. O'Connell. (1999). An inverse of Sanov's Theorem. *Stat. & Prob. Letters*, 42:201–206.

[7] Ghosal, A., Ghosh, J. K. and R. V. Ramamoorthi (1999). Consistency issues in Bayesian Nonanparametrics. *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri*, 639–667, Dekker.

[8] Grendár. M. (2005). Conditioning by rare sources. *Acta Univ. Belii, Math.* 12: 19-29. Online at http://actamath.savbb.sk.

[9] Grendár , M. (2006) L-divergence consistency for a discrete prior. *Jour. Stat. Res.* 40/1:73-76.

[10] Grendár, M. and G. Judge (2006). Large deviations theory and empirical estimator choice. *Econometric Rev.*, To appear.

[11] Grendár, M. and G. Judge (2006). Conditional Limit Theorems and Statistics. *Working paper.*

[12] Kitamura, Y. and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica.* 65:861-874.

[13] Kitamura, Y. and M. Stutzer (2002). Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics.* 107:159-174.

[14] Kleijn, B. J. K. and A. W. van der Vaart (2002). Misspecification in infinite-dimensional Bayesian statistics. *preprint.*

[15] Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika.* 90:319-326.

[16] Mittelhammer, R., Judge, G. and D. Miller (2000). *Econometric Foundations.* Cambridge:CUP.

[17] Owen, A. (2001). *Empirical Likelihood.* New York:Chapman-Hall/CRC.

[18] Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* 19:1725-1747.

[19] Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* 22:300-325.

[20] Ragusa, G. (2006). Bayesian likelihoods for moment condition models. *Working paper.* Univ. of California, Irvine.

[21] Rubin, D. (1981). Bayesian bootstrap. *Ann. Stat.* 9:130-134.

[22] Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika.* 92:31-46.

[23] Walker. S. (2004). New appraoches to Bayesian consistency. *Ann. Statist.* 32:2028–2043.

[24] Walker, S., Lijoi, A. and I. Prünster (2004). Contibutions to the understanding of Bayesian consistency. *working paper.*

Marian Grendár
Department of Mathematics
Faculty of Natural Sciences, Bel University
Tajovskeho 40
SK-974 01 Banska Bystrica
Slovakia
E-mail: marian.grendar@savba.sk

George Judge
Professor in the Graduate School
207 Gainnini Hall
Univesity of California, Berkeley
CA, 94720
E-mail: judge@are.berkeley.edu