Title

Consistency of Empirical Likelihood and Maximum A-Posteriori Probability Under Misspecification

Authors

Grendar, Marian
Judge, George G.

Publication Date

2008-02-20

CONSISTENCY OF EMPIRICAL LIKELIHOOD AND MAXIMUM A-POSTERIORI PROBABILITY UNDER MISSPECIFICATION

By Marian Grendár* and George Judge

Bel University and University of California, Berkeley

Using a large deviations approach, Maximum A-Posteriori Probability (MAP) and Empirical Likelihood (EL) are shown to possess, under misspecification, an exclusive property of Bayesian consistency. Under conditions of consistency, regardless of prior the MAP estimator asymptotically coincides with EL. The consistency property is also studied for sampling processes other than iid.

1. Introduction. Owen's [26], [27] Empirical Likelihood Theorem lays at the ground of Empirical Likelihood (EL) approach to inference, which is based on nonparametric likelihood ratio statistic. Building on Owen's result, Qin and Lawless [28] formulated EL with Estimating Equations (EE) and demonstrated under a basic set of regularity conditions, asymptotic normality of the resulting EL estimator. These two results turned EL into an attractive 'orthodox' semiparametric method of estimation and inference, that encapsulates the Maximum Nonparametric Likelihood method (MNPL) as a nonparametric special case.

Recently, Schennach [32] noted that the empirical likelihood function 'has not formally been shown to have a well-defined probabilistic interpretation that would justify its use in Bayesian inference'. The only convincing evidence supporting its Bayesian use comes from [22]. In [22] Lazar listed possible ways of turning EL into a Bayesian method and used the framework of Monahan and Boos [24] to assess by Monte Carlo simulations whether posterior density obtained from applying Bayes rule with empirical likelihood can be interpreted as the usual posterior obtained from model-based likelihood. Owen ([27], Chap. 9) notes a similarity between the EL and the Bayesian bootstrap [30], when in the latter, so-called non-informative Dirichlet prior distribution is assumed, however Sethuraman and Tivary [34] argue that such a prior is in fact highly informative. In order to provide an interpreta-

*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 nonparametric consistency, Bayesian Large Deviations, L-divergence, Pólya sampling, right censoring

tion of EL, Schennach [32] proposed a specific prior over a set of sampling distributions to get a Bayesian procedure that admits an operational form *similar* to EL. In [29] a different prior over the set of probability measures is considered and a group of EL-like methods is obtained.

In this paper, we use a large deviations approach to study Bayesian consistency under misspecification, in nonparametric and semiparametric models. The large deviations approach leads to the Bayesian Law of Large Numbers (BLLN) which means that in an infinite dimensional, possibly misspecified model Φ, the posterior weakly concentrates on L-projections of the 'true' sampling distribution on Φ. There are two methods which comply with BLLN and hence satisfy the consistency requirement: Bayesian Maximum A-Posteriori Probability (MAP) and 'orthodox' Maximum Non-Parametric Likelihood (MNPL), or in a semiparametric model, Empirical Likelihood (EL). Under the conditions of BLLN, regardless of prior, Bayesian MAP asymptotically turns into MNPL/EL. MAP and EL can be viewed, in terms of point estimation, as asymptotically identical methods. Consequently this sheds a new light on EL. The BLLN is demonstrated for independent, identically distributed (iid) data, multicolor Pólya sampling process, and right-censored data.

1.1. *Problem statement and theoretical base.* Given a sample of data, selecting a sampling distribution from a set Φ of such a distributions, is an important statistical problem. In this context Φ can be parametrized by a finite or infinite-dimensional parameter. The flexibility of the nonparametric (i.e., infinite-dimensional) model can be combined with advantages of finite parametrization. For instance, assume that a researcher is not willing to specify a parametric family $q(x;\theta)$ ($\theta \in \Theta \subseteq \mathbb{R}^m$, m finite) of data-sampling distributions, but is only willing to specify some of its underlying features. These features, i.e., the model $\Phi(\Theta)$, can be characterized by Estimating Equations (cf. [12], [23], Chap. 11): $\Phi(\Theta) \triangleq \bigcup_\Theta \Phi(\theta)$, where $\Phi(\theta) \triangleq \{q(x;\theta) : \int q(x;\theta)u_j(x;\theta) = 0, 1 \leq j \leq J\}$, $\theta \in \Theta \subseteq \mathbb{R}^K$, and J – the number of estimating functions $u(\cdot)$ – need not be equal to the number K of parameters. In general, a feasible set $\Phi(\Theta)$ of nonparametric sampling distributions which are indexed by a parameter θ, can be formed in a way other than by EE. The purely nonparametric Φ is contained in $\Phi(\Theta)$ as a special case. The problem of selection of sampling distribution from $\Phi(\Theta)$ can be considered also in Bayesian framework, where if a prior distribution Π over the set $\Phi(\Theta)$ is assumed, a prior distribution $\Pi(\theta)$ over Θ is induced.

In practice, the 'true' data sampling distribution $r(X;\theta)$ need not belong to the model set; i.e., the model can be misspecified. Assuming the Bayesian

framework, it is of interest to know the sampling distribution(s) on which the posterior measure concentrates, as n, the sample size gets large. Importance of the frequentist concept of consistency for Bayesian statistics can be justified both from subjectivist and objectivist Bayesian positions; cf. [38] or [11], Chap. 4.

We use a Large Deviations (LD) (cf. Ben-Tal, Brown and Smith [1], [2], Ganesh and O'Connell [9]) approach to Bayesian nonparametric consistency. The approach results in a Bayesian Sanov Theorem (BST) and its corollary the Bayesian Law of Large Numbers (BLLN) which establishes the consistency. LD theory is a sub-field of probability theory where, informally, the typical concern is about the asymptotic behavior, on a logarithmic scale, of the probability of a given event. The Bayesian Sanov Theorem identifies the rate function governing exponential decay of the posterior measure, and this in turn identifies the sampling distributions on which the posterior concentrates, as those distributions that minimize the rate function. Currently used approaches to Bayesian nonparametric consistency (cf. [37]) do not recognize this concentration of the posterior measure as a solution of the optimization problem.

The Bayesian Law of Large Numbers may be, informally, stated as follows: if the prior over a set Φ of sampling distributions, which might not include the 'true' distribution with pdf r, satisfies certain conditions, then the posterior asymptotically concentrates (a.s. r^∞) on weak neighborhoods of the L-projections of r on Φ. L-projection $\hat{q}$ of r on Φ is $\hat{q} = \arg\inf_{q\in\Phi} L(q\,||\,r)$, where $L(q\,||\,r)$ is the L-divergence of pdf q wrt r. In the case of iid sampling, $L(q\,||\,r) = -\int r \log q$.

If the conditions of the BLLN, called Schwartz conditions, are satisfied, then it can be shown (cf. Lemma 2.2, Sect 2.1) that distributions that maximize the nonparametric likelihood asymptotically (a.s. r^∞) turn into L-projections. Hence, selecting Maximum Non-Parametric Likelihood (MNPL) sampling distribution(s) complies with BLLN. In similar manner, it can be shown (cf. Lemma 2.1, Sect 2.1) that also distributions with the highest value of posterior probability asymptotically (a.s. r^∞) turn into L-projections. Hence, selection of Maximum A-posteriori Probable (MAP) distributions also satisfies the consistency requirement. To sum up, under misspecification MNPL and MAP are the only two methods with the Bayesian nonparametric consistency property. Selection of a posterior mean, or sampling distribution that minimizes say Kullback Leibler distance $I(q\,||\,r) = \int q \log \frac{q}{r}$ wrt q, in a misspecified case, would in general, violate the BLLN.

The BLLN directly applies also to 'semiparametric' models $\Phi(\Theta)$, where it achieves consistency under misspecification for MAP and the Empirical

Likelihood.

Lemmas 2.1 and 2.2, which are a by-product of the BLLN, make it possible to address in a new way the lack of probabilistic Bayesian interpretation, mentioned by Schennach. The lemmas demonstrate that EL (or MNPL) and MAP estimators asymptotically coincide.

In other words, if the Schwartz conditions are satisfied, then regardless of the prior used in Bayesian nonparametric model, the Bayesian MAP estimator asymptotically turns into MNPL estimator; or EL estimator in the semiparametric case. However, it should be added with the same breath that distributional asymptotic properties of MNPL and MAP might be different, since the Bernstein - von Mises theorem does not apply even for a simple infinite dimensional models; cf. Freedman [8].

1.2. *Organization of the paper.* In Sect. 2.1 formal framework is established and Bayesian nonparametric consistency is defined. There, also L-divergence is introduced and the BST and BLLN theorems are proven for the iid case. In Sect. 2.2, the BST and the BLLN for the semiparametric model, based on Estimating Equations, are discussed. It is shown there that the L-projection, singled out by the BLLN, is an asymptotic form of the EL and MAP estimators. In order to further explore consistency under misspecification and expand the scope of the related asymptotic connection between MNPL/EL and MAP, we prove the BLLN also for multicolor Pólya sampling process (Sect. 2.3) where using the BLLN suggests two possible variants of MNPL; and for right-censored data (Sect. 2.4), showing that Kaplan Meier estimator is asymptotic form of Bayesian MAP. In Section 3, a few open problem are briefly formulated.

2. Bayesian LLN's, MNPL, EL and MAP. The BLLN is a Corollary of a Sanov Theorem for Sampling Distributions which we call the bayesian Sanov Theorem (BST), for short. The BST is Bayesian counterpart of a Sanov Theorem for Empirical Measures (cf. [31], [3] Sect. III, VII and references cited therein). The latter, as well as its corollary, the Conditional Law of Large Numbers, are basic results of Large Deviations (LD) theory (cf. [3], [5]). The LD theorems for empirical measures have a bearing for the Relative Entropy Maximization method, and, as it was recognized in [19], also for its estimating equations extension: Exponential Tilting [15], [18], aka Maximum Entropy Empirical Likelihood [23]. In fact, the work [19] of Kitamura and Stutzer served as a starting point for our attempt to provide a similar underpinning to the MNPL and EL methods. It turned out, that this is only possible in a Bayesian framework.

2.1. *BLLN for iid sampling.* Let $\mathcal{P}$ be the set of all probability measures on $(\mathbb{R}, \mathcal{B})$, which are dominated by the Lebesgue measure. Let $X_1, X_2, \ldots$ be iid random variables, that take values in $(\mathbb{R}, \mathcal{B})$, with probability density function (pdf) r, where probability densities are denoted by lower case. $\mathcal{P}$ is endowed with weak topology. Let $\Phi \subseteq \mathcal{P}$. It is assumed that r, the true sampling distribution, is not necessarily in Φ. Let $\sigma(\mathcal{P})$ be a Borel σ-field on $\mathcal{P}$. A positive prior Π is put on $(\mathcal{P}, \sigma(\mathcal{P}))$ that is strictly positive over Φ. The prior combines with data $X_1^n \triangleq X_1, X_2, \ldots, X_n$ to define the posterior distribution

$$\Pi_n(Q|X_1^n) = \frac{\int_Q e^{-l_n(q)}\Pi(dq)}{\int_\Phi e^{-l_n(q)}\Pi(dq)},$$

where $l_n(q) \triangleq -\sum_{i=1}^n \log q(x_i)$ (log means the base e); $Q \subseteq \Phi$.

Let d be a metric on $\mathcal{P}$. The sequence $\{\Pi_n(\cdot|X_1^n), n \geq 1\}$ is said to be d-consistent at r, if there exists a $\Omega_0 \subset \mathbb{R}^\infty$ with $r(\Omega_0) = 1$ such that for $\omega \in \Omega_0$, for every neighborhood U of r, $\Pi_n(U|X_1^n) \to 1$ as n goes to infinity. If a posterior is d-consistent for any $r \in \Phi$ then it is said to be d-consistent. If the consistency holds for Hellinger distance, then the posterior is strongly consistent. If convergence holds in weak topology, the posterior is said to be weakly consistent. In [36] a decision-theoretic argument is proposed, in favor of weak consistency. Surveys of Bayesian nonparametric consistency can be found in [11], [10], [38].

To the best of our knowledge, Ben-Tal, Brown and Smith [1] were the first to use an LD approach to Bayesian nonparametric consistency. The authors showed consistency for X taking values from a finite set $\mathcal{X}$ and possibly misspecified model. Recently, independently Ganesh and O'Connell [9] established the first formal BST, for finite set $\mathcal{X}$ and a well-specified model. Here we develop the BST and the BLLN for $\mathcal{X} = \mathbb{R}$ and a possibly misspecified model. Using techniques other than LD, consistency in Hellinger distance and under misspecification was studied by Kleijn and van der Vaart [20].

The key quantity that governs the LD exponential decay of the posterior $\Pi_n(Q|X_1^n)$ is in the iid case the L-divergence of $q \in \mathcal{P}$ wrt $p \in \mathcal{P}$: $L(q\,||\,p) \triangleq -\int p \log q$; cf. [13]. In the discrete case L-divergence appears in Freedman's ([7], Thm 1) as 'entropy'. If p is an empirical pmf, then L-divergence appears as Kerridge's inaccuracy [17], [21], which is just the negative of the nonparametric likelihood. The L-projection $\hat{q}$ of p on $Q \subseteq \mathcal{P}$ is $\hat{q} \triangleq \arg\inf_{q \in Q} L(q||p)$. The value of L-divergence, at an L-projection of p on Q, is denoted by $L(Q||p)$.

Finally, let for $p, q \in \mathcal{P}$, $\epsilon > 0$, $B_\epsilon(q, p) \triangleq \{q' \in \mathcal{P} : L(q'\,||\,p) - L(q\,||\,p) < \epsilon\}$. For $A \subseteq \mathcal{P}$, $B_\epsilon(A, p) \triangleq \{q \in \mathcal{P} : L(q\,||\,p) - L(A\,||\,p) < \epsilon\}$.

Using this notation, the BST can be stated as follows:

THEOREM 2.1. (BST) *Let X_1^n be i.i.d. r. Let Q, Φ be open in weak topology; $Q \subset \Phi \subseteq \mathcal{P}$. Let $L(Q\,||\,r) < \infty$; for any $\epsilon > 0$, let $\Pi(B_\epsilon(Q,r)) > 0$ and $\Pi(B_\epsilon(\Phi,r)) > 0$. Then for $n \to \infty$,*

$$\frac{1}{n}\log \Pi_n(q \in Q|X_1^n) = -\{L(Q\,||\,r) - L(\Phi\,||\,r)\}, \qquad \text{a.s. } r^\infty.$$

PROOF. For $S \subseteq \mathcal{P}$, $l_n(S) \triangleq \inf_{q\in S} l_n(q)$. Let for $\epsilon > 0$, $B_\epsilon^n(S) \triangleq \{q : l_n(q) - l_n(S) < \epsilon\}$. Then $\int_A e^{-l_n(A)}\Pi(dq)$, $A = \{Q, \Phi\}$, can be bounded as:

$$e^{-l_n(A)-\epsilon}\Pi(B_\epsilon^n(A) \cap A) \leq \int_A e^{-l_n(q)}\Pi(dq) \leq e^{-l_n(A)}.$$

By lower-semicontinuity of L-divergence in weak topology and Strong Law of Large Numbers (which can be applied, since $L(Q\,||\,r) < \infty$, by assumption), $\frac{1}{n}l_n(A) \to L(A\,||\,r)$, a.s. r^∞, as $n \to \infty$. Thus, it holds: $\limsup_{n\to\infty} \frac{1}{n}\log \int_A e^{l_n(q)}\Pi(dq) \leq -L(A\,||\,r)$. So, $\limsup_{n\to\infty} \Pi_n(Q|X_1^n) \leq -\{L(Q\,||\,r) - L(\Phi\,||\,r)\}$. By the same argument (SLLN and continuity), for sufficiently large n, $\Pi(B_\epsilon^n(A)) > 0$, since $\Pi(B_\epsilon(A,r)) > 0$ by assumption. As $B_\epsilon^n(A) \cap A \neq \emptyset$, thus $\lim_{n\to\infty} \frac{1}{n}\log\, \Pi(B_\epsilon^n(A) \cap A) = 0$. Hence, $\liminf_{n\to\infty} \Pi_n(Q|X_1^n) \geq -\{L(Q\,||\,r) - L(\Phi\,||\,r)\}$. □

The posterior probability $\Pi_n(Q|X_1^n)$ decays exponentially fast with the decay rate $L(Q\,||\,r) - L(\Phi\,||\,r)$. The BST implies the Bayesian Law of Large Numbers (BLLN).

THEOREM 2.2. (BLLN) *Let $\Phi \subseteq \mathcal{P}$ be open in weak topology. Let 1) for every $q \in \Phi$, $\Pi(B_\epsilon(q,r)) > 0$, 2) $L(\Phi\,||\,r) < \infty$. Let $U \triangleq \bigcup_k W(\hat{q}_k, \epsilon)$, be union of weak ϵ-balls $W(\hat{q}_k, \epsilon)$ centered at L-projections $\hat{q}_k$, $k = 1, \ldots, \kappa$, $\kappa < \infty$, of r on Φ. Then,*

$$\lim_{n\to\infty} \Pi_n(q \in U|X_1^n) = 1, \qquad \text{a.s. } r^\infty.$$

PROOF. Let $Q \subset \Phi$ be open sets in weak topology. First, let Q be any set such that $\infty > L(Q\,||\,r) > L(\Phi\,||\,r)$. Then assumptions of the BST are satisfied, and the theorem implies that $\Pi_n(Q|X_1^n) \to 0$, a.s. r^∞, as $n \to \infty$. Note that $L(Q\,||\,r) = \infty$ for such Q that the L-projection $\hat{q}^Q$ of r on Q has support that is smaller than the support of r. However, for such a $\hat{q}^Q$, the posterior probability would be zero. The posterior thus concentrates on L-projections of r on Φ, provided that their support is not smaller than that of r. This is guaranteed by the assumption $L(\Phi\,||\,r) < \infty$. □

The BLLN Theorem is an extension of Schwartz' consistency theorem [33], to the case of a misspecified model. Assumptions 1 and 2 of the BLLN, called hereafter Schwartz conditions, reduce in the well-specified case to the Kullback Leibler support condition (cf. [33], [11], Thm. 4.4.2).

The next Lemma points out that the Bayesian Maximum A-Posteriori Probability (MAP), which selects $\hat{q}_{\text{MAP}} \triangleq \arg\sup_{q\in\Phi} \Pi_n(q \,|\, X_1^n)$ satisfies the BLLN.

LEMMA 2.1. *Let $\Phi \subseteq \mathcal{P}$ be open and Schwartz conditions be satisfied. Then, as $n \to \infty$, the set of* MAP *distributions $\mathcal{M} \triangleq \{\hat{q}_{\text{MAP}} : \hat{q}_{\text{MAP}} = \arg\sup_{q\in\Phi} \Pi_n(q \,|\, X_1^n)\}$ converges (a.s. r^∞) to the set of L-projections of r on Φ.*

PROOF. Thanks to the Strong LLN (SLLN), which can be applied under the Schwartz condition 2, conditions for infimum of minus the logarithm of the posterior probability (positivity of which is guaranteed by the Schwartz condition 1) turn into those for L-projections. □

Directly from the Strong LLN it follows that the Maximum Non-Parametric Likelihood (MNPL), that selects $\hat{q}_{\text{MNPL}} \triangleq \arg\inf_{q\in\Phi} l_n(q)$, satisfies the BLLN.

LEMMA 2.2. *Let $\Phi \subseteq \mathcal{P}$ be open and Schwartz condition 2 be satisfied. Then, as $n \to \infty$, the set of* MNPL *distributions converges (a.s. r^∞) to the set of L-projections of r on Φ.*

The lemmas also mean that the MNPL and the MAP methods asymptotically select the same sampling distribution(s).

Next, we turn to semiparametric setting.

2.2. *BLLN for the semiparametric $\Phi(\Theta)$.* Let X be random variable with pdf $r(X;\theta)$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^K$. A Bayesian specifies a model $\Phi(\Theta)$ (cf. Sect. 1.1) and puts a positive prior Π over $\Phi(\Theta)$, which in turn, induces a prior $\Pi(\theta)$ over Θ; see Florens and Rolin [6], where also several models are worked out, using a Dirichlet process prior. If requirements of the BLLN are satisfied, then the posterior $\Pi_n(\cdot|X_1^n)$ concentrates on weak neighborhoods of L-projections $\hat{q}$ of r on $\Phi(\Theta)$

$$\hat{q}(x;\hat{\theta}) = \arg \inf_{q(x;\theta)\in\Phi(\theta)} \inf_{\theta\in\Theta} L(r \,||\, q(x;\theta)).$$

The most common form of $\Phi(\Theta)$ is the one defined by Estimating Equations (cf. Sect 1.1). In this case $\Phi(\Theta)$ is also known as a linear family of distributions, that we denote as $\mathcal{L}(u)$. The L-projection of r on $\mathcal{L}(u)$ can be found

by means of the following Theorem 2.3. To state it, we introduce a Λ family of distributions and recall a concept of support of a convex set. Let Λ be a family of pdf's: $\Lambda(r,u,\lambda,\theta) \triangleq \{p \in \mathcal{P} : p = r[1-\sum_{j=1}^{J} \lambda_j u_j(\cdot;\theta)]^{-1}, \lambda \in \mathbb{R}^J\}$. The support $S(\mathcal{C})$ of a convex set $\mathcal{C} \subset \mathcal{P}$ is just the support of the member of $\mathcal{C}$ for which $S(\cdot)$ contains the support of any other member of the set.

THEOREM 2.3. *Let $\Phi = \mathcal{L}(u)$. Let $r \in \mathcal{P}$ be such that $S(r) = S(\mathcal{L})$. Then the L-projection $\hat{q}$ of r on Φ is unique and belongs to the $\Lambda(r,u,\lambda,\theta)$ family; i.e., $\mathcal{L}(u) \cap \Lambda(r,u,\lambda,\theta) = \{\hat{q}\}$.*

PROOF. In light of Theorem 9 of [4] it suffices to check that $\hat{q} = r[1 - \sum_{j=1}^{J} \lambda_j u_j(\cdot;\theta)]^{-1}$, with λ such that $\hat{q} \in \mathcal{L}(u)$, satisfies: $\int_{S(r)} r \left(1 - \frac{q'}{\hat{q}}\right) = 0$, for all $q' \in \Phi$, which is indeed the case. □

The estimator $\hat{\theta}$ can, thanks to convex duality, be obtained as

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \sup_{\lambda \in \mathbb{R}^J} L(r \,||\, q(x;\lambda,\theta)),$$

where $q(x;\lambda,\theta) \in \Lambda(r,u,\lambda,\theta)$. Since r is in practice not known, Kitamura and Stutzer [18] suggested that $L(r \,||\, q(x;\lambda,\theta))$ be replaced by its estimate $\hat{L}(q(x;\lambda,\theta)) \triangleq -\frac{1}{n}\sum_{i=1}^{n} \log q(x_i;\lambda,\theta)$. The resulting estimator

$$\hat{\theta}_{\mathrm{EL}} \triangleq \arg \inf_{\theta \in \Theta} \sup_{\lambda \in \mathbb{R}^J} \hat{L}(q(x;\lambda,\theta)) \tag{2.1}$$

is just the Empirical Likelihood (EL) estimator [28], [27], since (2.1) is a convex dual problem to the following optimization problem, by means of which EL is usually defined: $\hat{q}_{\mathrm{EL}}(x;\hat{\theta}_{\mathrm{EL}}) = \arg \sup_{q(x;\theta)\in\Phi(\theta)} \sup_{\theta\in\Theta} \sum_{i=1}^{n} \log q(x_i;\theta)$. Analogous to Lemma 2.2 it can be shown that the EL estimator $\hat{q}_{\mathrm{EL}}(x;\hat{\theta}_{\mathrm{EL}})$ asymptotically (a.s. r^∞) turns into an L-projection of r on $\Phi(\Theta)$. The same holds for the MAP estimator

$$\hat{q}_{\mathrm{MAP}}(x;\hat{\theta}_{\mathrm{MAP}}) = \arg \sup_{q(x;\theta)\in\Phi(\theta)} \sup_{\theta\in\Theta} \Pi_n(q(x;\theta)|x_1^n).$$

Hence, the EL and the MAP estimators are consistent under misspecification. This provides a basis for the EL approach as well for the Bayesian MAP estimation.

In the univariate case, BST and BLLN (Thm 2.1, 2.2), X can be replaced by a multivariate random variable and the theorems remain valid. Consequently, the extension to Φ, constructed by multivariate EE, is also direct. As an example, consider the linear regression model $Y = \alpha + \beta X + \epsilon$, with

stochastic X. In EE this is usually approached through estimating equations: $\Phi(\theta) = \{q(x,y;\theta) : \int q(x,y;\theta)[Y-(\alpha+\beta X)] = 0, \int q(x,y;\theta)X[Y-(\alpha+\beta X)] = 0\}$, $\theta \triangleq (\alpha,\beta) \in \mathbb{R}^2 \equiv \Theta$, which are based on the gaussian model score equations. The multivariate BLLN shows that the posterior asymptotically concentrates on the L-projections of r on $\Phi = \bigcup_\Theta \Phi(\theta)$. Again, there are two methods which comply with the BLLN: EL and MAP.

2.3. *BLLN for Pólya sampling.* In this section we prove the BST and the BLLN for multi-color Pólya urn - a simple sampling process where data are neither identically nor independently distributed. The theorems can be directly used also in a corresponding semiparametric $\Phi(\Theta)$ setting.

The probability of a sample X_1^n being drawn from a multicolor Pólya urn with parameter $c \in \mathbb{Z}$ and initial configuration $q(N) \triangleq (\alpha_1, \dots \alpha_m)/N$, is $\log \Pi(X_1^n \,|\, q(N); c) \triangleq \sum_{i=1}^m \sum_{l=0}^{n_i-1} [\log(\alpha_i + jc) - \log(N+jc)]$; this is meaningful if $-n\,c \le \min(\alpha_1, \dots, \alpha_m)$. We embed the sampling scheme into a Bayesian nonparametric setting. To this end, let $\mathcal{P}(\mathcal{X})$ be set of all probability mass functions with the support $\mathcal{X} = \{x_1, \dots, x_m\}$. Let $\Phi \subseteq \mathcal{P}(\mathcal{X})$ and let $\Phi(N)$ denote intersection of Φ with the set of all possible configurations of the N-urn. Let $\Phi(N)$ be the support of the prior distribution $\Pi(q(N))$ of initial configurations $q(N)$. Let $r(N)$ be the true initial configuration, where $r(N)$ is not necessarily in $\Phi(N)$. As before, we are interested in the LD asymptotics of the posterior distribution $\Pi_n(q(N) \,|\, X_1^n; c)$. Asymptotic investigations of posterior consistency will be carried on under the following assumptions: 1) n and N goes to infinity in such a way that $\beta(n) \triangleq \frac{n}{N} \to \beta \in (0,1)$ as $n \to \infty$, 2), $r(N)$ converges in the total variation metrics to $r \in \mathcal{P}(\mathcal{X})$ as $n \to \infty$. Topological qualifiers are meant in topology induced on the m-dimensional simplex by the usual topology on $\mathbb{R}^m$.

The exponential decay of posterior is governed by Pólya L-divergence. For $p, q \in \mathcal{P}(\mathcal{X})$, the Pólya L-divergence $L_\beta^c(q \,||\, p)$ of q with respect to p is

$$L_\beta^c(q \,||\, p) \triangleq -\sum_{i=1}^m p_i \log(q_i + \beta c p_i) + \frac{1}{\beta c} \sum_{i=1}^m q_i \log \frac{q_i}{q_i + \beta c p_i}.$$

By the continuity argument, $L_\beta^0(q \,||\, p) \triangleq -\sum_{i=1}^m p_i \log q_i - 1$. The Pólya L_β^c-projection $\hat{q}$ of p on $Q \subseteq \mathcal{P}(\mathcal{X})$ is $\hat{q} \triangleq \arg\inf_{q \in Q} L_\beta^c(q \,||\, p)$. The value of L_β^c-divergence at an L_β^c-projection of p on Q is denoted by $L_\beta^c(Q||p)$.

THEOREM 2.4. *Let $Q \subset \Phi$ be an open set. Let $\beta(n) \to \beta \in (0,1)$, $r(N) \to r$ as $n \to \infty$. Let $L_\beta^c(Q||r) < \infty$. Then, for $n \to \infty$,*

$$\frac{1}{n} \log \Pi_n(q(N) \in Q \,|\, X_1^n; c) = -\left\{ L_\beta^c(Q \,||\, r) - L_\beta^c(\Phi \,||\, r) \right\}$$

with probability one.

PROOF. The proof is constructed separately for $c > 0$, $c < 0$, $c = 0$.

For $c \neq 0 \wedge \eta c \notin (\mathbb{Z}^-)^m \wedge \eta \notin \mathbb{Z}^-$, $\log \Pi(X_1^n \,|\, q(N); c)$ can equivalently be expressed as $\log(\Gamma(\eta)/\Gamma(\eta+n)) + \sum_{i=1}^m \log(\Gamma(\eta q_i + n_i)/\Gamma(\eta q_i))$, where $\Gamma(\cdot)$ is the Gamma function and $\eta \triangleq N/c$. For $0 < a < b$, the ratio $\Gamma(b)/\Gamma(a)$ can be upper-bounded by $b^{b-1/2}/a^{a-1/2}e^{b-a}$ and lower-bounded by $b^{b-1}/a^{a-1}e^{b-a}$, cf. [16]. Then, $\Pi_n(q(N) \in Q \,|\, x^n; c)$ can be upper-bounded by U_n (dependence of q on N is made implicit):

$$U_n = \frac{\sum_{q \in Q} \Pi(q) \prod_{i=1}^m e^{-n\, l(q_i, \frac{1}{2n})}}{\sum_{q \in \Phi} \Pi(q) \prod_{i=1}^m e^{-n\, l(q_i, \frac{1}{n})}}$$

and lower-bounded by L_n in similar way; to get L_n just replace $1/2n$ with $1/n$ in U_n. There, $l(q_i, \alpha) \triangleq -[(\gamma_i - \alpha)\log\gamma_i + (\gamma_i + \nu_i^n - \alpha)\log(\gamma_i + \nu_i^n)]$, $\gamma_i \triangleq \frac{q_i}{\beta(n)c}$, $\alpha \in \left\{\frac{1}{n}, \frac{1}{2n}\right\}$ and ν^n is the empirical measure induced by the sample X_1^n.

Next, we use simple bounds to upper bound U_n by $\bar{U}_n$

$$\bar{U}_n = \frac{\prod_{i=1}^m e^{-n\, l\left(\hat{q}_i(Q, \frac{1}{2n}), \frac{1}{2n}\right)}}{\pi(\hat{q}(\Phi, \frac{1}{n})) \prod_{i=1}^m e^{-n\, l\left(\hat{q}_i(\Phi, \frac{1}{n}), \frac{1}{n}\right)}},$$

and to lower bound L_n by $\underline{L}_n$; to get $\underline{L}_n$ just replace $1/2n$ with $1/n$ in $\bar{U}_n$. There, $\hat{q}(\cdot, \alpha) \triangleq \arg\inf_{q \in \cdot} \sum_{i=1}^m l(q_i, \alpha)$.

By the Strong Law of Large Numbers for Pólya Sampling, $\nu^n \to r$, almost surely, as $n \to \infty$. The Pólya L-divergence is continuous in q and Q is open, by assumption. Thus, $\frac{1}{n}\log\bar{U}_n$ converges, with probability one, to $-\left\{L_\beta^c(Q \,||\, r) - L_\beta^c(\Phi \,||\, r)\right\}$, as $n \to \infty$. This is the same as the 'point' of almost sure convergence of $\frac{1}{n}\log\underline{\mathrm{L}}_n$ and the Theorem for $c > 0$ is thus proven.

For $c \neq 0 \wedge (1 - \eta q) \notin (\mathbb{Z}^-)^m \wedge (1 - \eta) \notin \mathbb{Z}^-$, $\log \Pi(X_1^n \,|\, q(N); c)$ can equivalently be expressed as $\log(\Gamma(1 - \eta - n)/\Gamma(1 - \eta)) + \sum_{i=1}^m \log(\Gamma(1 - \eta q_i)/\Gamma(1 - \eta q_i - n_i))$. The proof then can be constructed along the same lines as for $c > 0$. The case of $c = 0$ is straightforward. □

From the Pólya BST (Thm 2.5), the BLLN for Pólya sampling directly follows. It is worth noting that the MNPL in Pólya sampling can be constructed in two ways: either via maximization of $\Pi(X_1^n \,|\, q(N); c)$, or by maximization of negative of $L_\beta^c(q||\nu^n)$ wrt q, where ν^n is empirical pmf induced by sample X_1^n. The methods could be called 'exact' and 'asymptotic' MNPL, respectively. Both the methods comply with Pólya BLLN, as does the Bayesian MAP.

2.4. *BST for right-censored data.* Right-censoring of a r.v. X by a r.v. Y (both on $(\mathbb{R}, \mathcal{B})$) can be described by the following hierarchical model: $\delta \sim \text{Ber}(\alpha)$, $\alpha \triangleq \int F_0(y)\, dG_0(y)$; if $\delta = 0$ then $X \sim F_0$; if $\delta = 1$ then $X = (Y, \infty)$ where $Y \sim G_0$; X's are conditionally independent. A Bayesian puts positive prior over the set Φ of distributions of X. Let the prior over distributions of Y be without loss of generality concentrated at G_0. We are interested in the exponential decay of the posterior

$$\Pi_n(F \in Q \,|\, X_1^n) = \frac{\int_Q e^{-l_n(F,n_1)} \Pi(dF)}{\int_\Phi e^{-l_n(F,n_1)} \Pi(dF)},$$

where $l_n(F, n_1) \triangleq -\sum_{i:\delta_i=0} \log F(\{X_i\}) - \sum_{i:\delta_i=1} \log F((Y_i, \infty))$, $Q \subset \Phi$, F_0 is not necessarily in Φ, and n_1 is the number of non-censored data, out of n observations. The decay is governed by the L-divergence of F wrt (F_0, G_0) for right-censoring

$$L(F \,||\, (F_0, G_0)) \triangleq -\left[\alpha \int \log F(x)\, dF_0(x) + (1-\alpha) \int \log F((y,\infty))\, dG_0(y)\right].$$

The L-projection $\hat{F}$ of (F_0, G_0) on $Q \subseteq \mathcal{P}$ is $\hat{F} \triangleq \arg\inf_{F \in Q} L(F \,||\, (F_0, G_0))$, and $L(Q \,||\, (F_0, G_0))$ denotes value of the L-divergence at an L-projection of F_0 on Q. Let $B_\epsilon(Q, F_0) \triangleq \{F \in \mathcal{P} : L(F \,||\, (F_0, G_0)) - L(Q \,||\, (F_0, G_0)) < \epsilon\}$. BST for right-censoring:

THEOREM 2.5. *Let X_1^n be right-censored data generated by the above model. Let Q, Φ be open in weak topology; $Q \subset \Phi \subseteq \mathcal{P}$. Let $L(Q \,||\, (F_0, G_0)) < \infty$, and for any $\epsilon > 0$, let $\Pi(B_\epsilon(Q, F_0)) > 0$, $\Pi(B_\epsilon(\Phi, F_0)) > 0$. Then for $n \to \infty$,*

$$\frac{1}{n} \log \Pi_n(F \in Q \,|\, X_1^n) = -\{L(Q \,||\, (F_0, G_0)) - L(\Phi \,||\, (F_0, G_0))\}.$$

PROOF. Note that $\frac{1}{n} l_n(F, n_1)$ converges to $L(F \,||\, (F_0, G_0))$, with probability 1, by the SLLN. Arguments go along the lines of the proof of Theorem 2.1. □

From the BST (Thm 2.6), the BLLN follows for right-censored data, in the same way as it does for iid case from Thm 2.1. The BLLN for right censoring demonstrates that posterior concentrates on weak neighborhoods of the L-projections of (F_0, G_0) on Φ, if the ϵ-balls $B_\epsilon(F, (F_0, G_0)) \triangleq \{F' \in \mathcal{P} : L(F' \,||\, (F_0, G_0)) - L(F \,||\, (F_0, G_0)) < \epsilon\}$, have positive prior probability. This together with assumption $L(\Phi \,||\, (F_0, G_0))$, form the Schwartz conditions for right-censoring.

Under the Shwartz conditions, a set of Bayesian MAP estimators $\hat{F}_{\mathrm{MAP}} \triangleq \arg\sup_{F\in\Phi} \Pi_n(F \mid X_1^n)$ asymptotically coincides with a set of L-projections of (F_0, G_0) on Φ. The same holds true for MNPL/EL estimator $\hat{F}_{\mathrm{EL}} \triangleq \arg\inf_{F\in\Phi} l_n(F, n_1)$. The Kaplan Meier estimator follows from $\hat{F}_{\mathrm{EL}}$ in the standard way, cf. [27]. Thus the BLLN makes it possible to view the Kaplan Meier estimator as an asymptotic instance of the Bayesian MAP, and provides a probabilistic underpinning. The only available Bayesian view of the Kaplan Meier estimator seems to be that of Susarla and van Ryzin [35]. In [35], a Dirichlet process prior was considered in a well-specified model, and it was shown there that the posterior mean converges to the Kaplan Meier estimator as the parameter α of the Dirichlet process converges to 0.

3. Open problems. The distributional properties of the MNPL/EL and MAP estimators in the misspecified model remain an open question. Also, the development of non-informative priors in a nonparametric context would be of interest and some proposals along this line can be found in [11]. The MNPL/EL as well as MAP satisfy Bayesian consistency under the misspecification requirement. There is also an un-conditional ('frequentist') counterpart of the requirement, which the Exponential Tilting method satisfies; cf. [19] and introduction of Sect. 2. The problem of selecting between the two requirements, and hence between the associated methods, was suggested in [14].

Acknowledgement. We are grateful to Nicole Lazar who gave us valuable feedback that encouraged us to write this note. Major impetus for improvement of previous version of this work came from reviewers, associate editor and main editor. Robert Niven's [25] opened possibility space to Sect. 2.4. We also want to thank Douglas Miller, Art Owen, Dale Poirier, Jing Qin and Giuseppe Ragusa for valuable comments and suggestions.

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] Csiszár I. (1998). The method of types. *IEEE Trans. Inform. Theory* 44:2505-2523
[4] Csiszár I. and P. Shields. (2004). Notes on Information Theory and Statistics: A tutorial. *Found. Trends Comm. Inform. Theory*, 1:1–111.
[5] Dembo, A. and O. Zeitouni. (1998). *Large Deviations Techniques and Applications*. New York:Springer-Verlag.
[6] Florens, J.-P. and J.-M. Rolin. (1994). Bayes, bootstrap, moments. Discussion paper 94.13. Institute de Statistique, Université catholique de Louvain.

[7] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* 34:1386-1403.
[8] Freedman, D. A. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* 27:1119-1140.
[9] Ganesh, A. and N. O'Connell. (1999). An inverse of Sanov's Theorem. *Statist. Probab. Lett.*, 42:201–206.
[10] 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.
[11] Ghosh, J. K. and R. V. Ramamoorthi. (2003). *Bayesian Nonparametrics*. New York: Springer.
[12] Godambe, V. P. and B. K. Kale. (1991). Estimating functions: an overview. In V. P. Godambe (ed.), *Estimating Functions*. Oxford, UK: Oxford University Press, pp. 3-20.
[13] Grendár. M. (2005). Conditioning by rare sources. *Acta Univ. M. Belii Ser. Math.* 12: 19-29. Online at http://actamath.savbb.sk.
[14] Grendár, M. and G. Judge. (2006). Large deviations theory and empirical estimator choice. *Econometric Rev.*, To appear.
[15] Imbens, G., R. Spady and P. Johnson. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* 66:333-357.
[16] Kečkić, J. D. and P. M. Vasić. (1971). Some inequalities for the gamma function. *Publ. Inst. Math. (Beograd)(N.S.)* 11:107-114.
[17] Kerridge, D. F. (1961). Inaccuracy and inference. *J. Roy. Statist. Soc. Ser. B.* 23:284-294.
[18] Kitamura, Y. and M. Stutzer. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica.* 65:861-874.
[19] Kitamura, Y. and M. Stutzer. (2002). Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics.* 107:159-174.
[20] Kleijn, B. J. K. and A. W. van der Vaart. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* 34(2):837-877.
[21] Kulhavý, R. (1996). *Recursive Nonlinear Estimation: A Geometric Approach*. Lecture Notes in Control and Information Sciences, vol. 216. London: Springer-Verlag.
[22] Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*. 90:319-326.
[23] Mittelhammer, R., Judge, G. and D. Miller. (2000). *Econometric Foundations*. Cambridge:CUP.
[24] Monahan, J. F. and D. D. Boos. (1992). Proper likelihoods for Bayesian analysis. *Biometrika* 79:271-278.
[25] Niven, R. K. (2005). Combinatorial information theory: I. philosophical basis of cross-entropy and entropy. On-line at arXiv:cond-mat/0512017.
[26] Owen, A. (1988). Empirical likelihood ratio confidence interval for a single functional. *Biometrika* 75(2):237-249.
[27] Owen, A. (2001). *Empirical Likelihood*. New York:Chapman-Hall/CRC.
[28] Qin, J. and J. Lawless. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* 22:300-325.
[29] Ragusa, G. (2006). Bayesian likelihoods for moment condition models. *Working paper*. Univ. of California, Irvine.
[30] Rubin, D. (1981). Bayesian bootstrap. *Ann. Statist.* 9:130-134.
[31] Sanov, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sb.* 42:11-44. In Russian.
[32] Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*.

92:31-46.
[33] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*. 4:10-26.
[34] Sethuraman, J. and R. C. Tiwari. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical decision theory and related topics, III, Vol. 2*, 305-315, Academic Press, New York,
[35] Susarla, V. and J. Van Ryzin. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* 71(356):897-902.
[36] Walker. S. and P. Damien. (2000). Practical Bayesian asymptotics. Working paper 00-007, Business School, Univ. of Michigan.
[37] Walker. S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* 32:2028–2043.
[38] Walker, S., Lijoi, A. and I. Prünster. (2004). Contibutions to the understanding of Bayesian consistency. Discussion paper 13/2004. ICER, Applied mathematics series.

This work is a follow up in a series of works, on which we participated, and which were not cited in the above text. Some aspects of these works will be explored further elsewhere. For an interested reader we provide a pointer to these works here:

[39] Grendár , M. (2006) L-divergence consistency for a discrete prior. *Jour. Stat. Res.* 40/1:73-76. Corrected at arXiv.
[40] Grendár, M. and G. Judge. (2006). Conditional Limit Theorems and Statistics. *Working paper*.
[41] Grendár, M. and G. Judge. (2007). Large Deviations Probabilistic Interpretation and Justification of Empirical Likelihood. *Preprint*, Department of Agricultural & Resource Economics, UCB. CUDARE Working Paper 1035, Apr. 2007. http://repositories.cdlib.org/are_ucb/1035
[42] Grendár, M. (2007). Maximum Probability and Maximum Entropy, Bayesian Maximum Probability and Maximum Non-parametric Likelihood. (Probabilistic Regularization of Inverse Problems). *unpublished qualification work*, June 2007.
[43] Grendár, M., Judge. G. and R. K. Niven. (2007). Large Deviations approach to Bayesian nonparametric consistency: the case of Polya urn. *Preprint*, Department of Agricultural & Resource Economics, UCB. CUDARE Working Paper 1048, Sep. 2007. http://repositories.cdlib.org/are_ucb/1048
[44] Grendár, M. (2008). Maximum Probability and Relative Entropy Maximization. Bayesian Maximum Probability and Empirical Likelihood. *Proc. of Intnl. Workshop on Applied Probability '08*; Feb 2008, accepted.

Marian Grendár
Department of Mathematics
Faculty of Natural Sciences, Bel University
Tajovskeho 40
SK-974 01 Banska Bystrica
Slovakia
Institute of Mathematics and CS
Slovak Academy of Sciences (SAS)
Institute of Measurement Sciences of SAS
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