Maximum Probability and Relative Entropy Maximization.

Bayesian Maximum Probability and Empirical Likelihood

M. Grendár

IWAP 2008

Contents

1. Φ-problem,
   Bayesian Maximum Probability,
   Maximum Non-Parametric Likelihood, and
   Empirical Likelihood
2. Π-problem,
   Maximum Probability,
   Relative Entropy Maximization, and
3. Empirical MaxMaxEnt

Within the two classes of problems, probabilistic justification and interpretation of the respective methods will be discussed.

The Φ-problem

There is

- a strictly positive prior distribution $\pi(\cdot)$
- over a nonparametric (i.e., infinite-dimensional) set Φ
- of data-sampling distributions and
- a sample $X_1, \ldots, X_n$
- from unknown 'true' data-sampling distribution r.

The objective is to select a data-sampling distribution from the set Φ, called model.

Consistency under misspecification requirement

As a rule, the model Φ in practice does not contain the 'true' data-sampling distribution r. The model is misspecified.

Requirement: a method for solving the Φ-problem should be such that it is consistent under misspecification; i.e., as n gets large, the method should select those sampling distributions on which the posterior distribution

$$\pi_n(Q \mid X_1^n) = \frac{\int_Q \{\prod_{i=1}^n q(x_i)\}\pi(dq)}{\int_\Phi \{\prod_{i=1}^n q(x_i)\}\pi(dq)}$$

concentrates.

Large Deviations approach to consistency under misspecification

Let $\mathcal{P}$ be the set of all probability mass functions (pmf's) with finite m-element support $\mathcal{X}$. The set $\mathcal{P}$ is endowed with the usual topology. Let $\Phi \subseteq \mathcal{P}$.

Large Deviations (LD): a study of the asymptotic behavior, on a logarithmic scale, of the probability of a given event. Bayesian Sanov Theorem (BST) 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.

L-divergence, L-projection

The L-divergence $L(q||p)$ of $q \in \mathcal{P}$ with respect to $p \in \mathcal{P}$:

$$L(q \,||\, p) \triangleq -\sum_{\mathcal{X}} p \log q.$$

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 $L(Q||p)$.

Bayesian Sanov Theorem

Theorem

(BST) *Let* X_1^n *be i.i.d.* r. *Let* $Q \subset \Phi \subseteq \mathcal{P}$; $L(Q \,||\, r) < \infty$.
Then for $n \to \infty$,

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

Bayesian Law of Large Numbers

Theorem

(BLLN) *Let $\Phi \subseteq \mathcal{P}$ be a convex, closed set.*
Let $B(\hat{q}, \epsilon)$, be a closed ϵ-ball defined by the total variation metric, centered at the L-projection $\hat{q}$ of r on Φ. Then,

$$\lim_{n \to \infty} \pi_n(q \in B(\hat{q}, \epsilon) \,|\, X_1^n) = 1, \quad a.s.\ r^\infty.$$

The posterior probability concentrates (a.s. r^∞) on the L-projection of the 'true' sampling distribution r on Φ.

The Bayesian Maximum Probability Method

Bayesian Maximum Probability method selects the Maximum A-Posteriori Probable (MAP) data-sampling distribution(s)

$$\hat{q}_{\mathrm{MAP}} \triangleq \arg\sup_{q\in\Phi} \pi(q) \prod_{i=1}^{n} q(x_i).$$

MAP satisfies the BLLN (a direct consequence of SLLN). The L-projection can be viewed as an asymptotic instance of MAP distribution.

Maximum Nonparameteric Likelihood

Also, Maximum Non-parametric Likelihood (MNPL) is consistent under misspecification. MNPL selects

$$\hat{q}_{\mathrm{MNPL}} \triangleq \arg\sup_{q \in \Phi} \sum_{i=1}^{n} \log q(x_i).$$

Summary of the Φ-problem

MNPL and MAP are consistent under misspecification. Selection of say Relative Entropy Maximizing sampling distribution or the posterior mean is an inconsistent method.

Parametric Φ-problem

Let X be a random variable with pmf $r(x;\theta)$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^K$. The model Φ, can be characterized by Estimating Equations (EE): $\Phi \triangleq \bigcup_{\Theta} \Phi(\theta)$, where

$$\Phi(\theta) \triangleq \left\{ q(x;\theta) : \sum_{\mathcal{X}} q(x;\theta) u_j(x;\theta) = 0, 1 \leq j \leq J \right\},$$

$u(\cdot)$ are estimating functions and $\theta \in \Theta \subseteq \mathbb{R}^K$. The number J of EE's may be greater than the number K of parameters.

The BLLN in parametric case

A Bayesian puts positive prior π over Φ, which in turn induces prior $\pi(\theta)$ over Θ.

By the BLLN, the posterior $\pi_n(\cdot|X_1^n)$ concentrates on a weak neighborhood of the L-projection $\hat{q}$ of $r(x;\theta)$ on Φ:

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

This thus provides a probabilistic justification for using $\hat{\theta}$ as an estimator of θ.

Parametric Φ-problem: Empirical Likelihood

Thanks to the convex duality, the estimator $\hat{\theta}$ can be obtained also as

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \inf_{\lambda \in \mathbb{R}^J} - \sum_{i=1}^{m} r(x_i) \log \left(1 - \sum_j \lambda_j u_j(x_i; \theta) \right) .$$

Since r is in practice not known, one can estimate the convex dual objective function by

$$- \sum_{l=1}^{n} \log \left(1 - \sum_j \lambda_j u_j(x_l; \theta) \right) .$$

The resulting estimator is the Empirical Likelihood (EL) estimator

The Π-problem

There is

- a known sampling distribution q
- a set $\Pi \subseteq \mathcal{P}$, into which an unavailable empirical pmf ν^n induced by an unavailable sample X_1^n, drawn from q, is assumed to belong.

The objective is to select an empirical pmf (aka type) from the set Π, called model.

Consistency under misspecification requirement

As a rule, the model Π in practice does not contain the 'true' data-sampling distribution r. The model is misspecified.

Requirement: a method for solving the Π-problem should be such that it is consistent under misspecification; i.e., as n gets large, the method should select those types which are conditionally (upon the event $\nu^n \in \Pi$) possible to observe.

Consistency under misspecification requirement (cont'd)

What is the set $B \subset \Pi$ such that, as $n \longrightarrow \infty$, the conditional probability

$$\pi(\nu^n \in B \,|\, \nu^n \in \Pi; q) \longrightarrow 1?$$

I-divergence, I-projection

I-divergence

$$I(p \,||\, q) \triangleq \sum p \log \frac{p}{q},$$

where $p, q \in \mathcal{P}$.

The I-projection $\hat{p}$ of q on $\Pi \subseteq \mathcal{P}$ is

$$\hat{p} \triangleq \arg \inf_{p \in \Pi} I(p \,||\, q).$$

The value of the I-divergence at an I-projection of q on Π is denoted by $I(\Pi||q)$.

Sanov Theorem

Theorem

(ST) *Let* Π *be an open set;* $I(\Pi \,||\, q) < \infty$.
Then, for $n \to \infty$,

$$\frac{1}{n} \log \pi(\nu^n \in \Pi; q) = -I(\Pi \,||\, q).$$

Conditional Law of Large Numbers

Theorem

(CLLN) *Let* Π *be a convex, closed set that does not contain* q*. Let* $B(\hat{p}, \epsilon)$ *be a closed* ϵ*-ball defined by the total variation metric that is centered at the I-projection* $\hat{p}$ *of* q *on* Π*. Then,*

$$\lim_{n \to \infty} \pi(\nu^n \in B(\hat{p}, \epsilon) \,|\, \nu^n \in \Pi; q) = 1.$$

Given that a type from Π was observed, it is asymptotically zero-probable that the type was different than the I-projection of the sampling distribution q on Π.

Maximum Probability Method

Maximum Probability (MaxProb) method selects the type

$$\hat{\nu}^n_{\mathrm{MaxProb}} = \arg \sup_{\nu^n \in \Pi} \pi(\nu^n; q)$$

which can be generated by the sampling distribution q, with the highest probability. If the sampling is i.i.d., then

$$\pi(\nu^n; q) = n! \prod_{i=1}^{m} \frac{q_i^{n_i}}{n_i!}.$$

It is easy to see that MaxProb satisfies the CLLN. The I-projections can be viewed as asymptotic instances of MaxProb types.

Relative Entropy Maximization Method

Relative Entropy Maximization method (REM, MaxEnt) selects the type with the highest value of relative entropy, i.e.,

$$\hat{\nu}^n_{\mathrm{REM}} = \arg \sup_{\nu^n \in \Pi} -\sum_{i=1}^{m} \nu^n \log \frac{\nu^n}{q}.$$

REM satisfies the CLLN.

Conventional REM/MaxEnt uses $-\sum p \log \frac{p}{q}$, $p \in \mathcal{P}$ as the objective function and this way it ignores the information about sample size n.

Summary of the Π-problem

MaxProb and REM are consistent under misspecification.
Selection of say MNPL sampling distribution or Rényi-Tsallis entropy maximizing type is an inconsistent method.

Parametric Π-problem

This time, the EE define a feasible set Π into which an unobserved parametrized type $\nu^n(\theta)$ is supposed to belong: $\Pi \triangleq \bigcup_\Theta \Pi(\theta)$, where

$$\Pi(\theta) \triangleq \left\{ p(x;\theta) : \sum_{\mathcal{X}} p(x;\theta) u_j(x;\theta) = 0, 1 \leq j \leq J \right\},$$

and $\theta \in \Theta \subseteq \mathbb{R}^K$. The objective is now to select parametric type $\nu^n(\theta)$ from Π.

The CLLN in parametric case

The CLLN implies that the parametric Π-problem should be (for $n \to \infty$) solved by selecting

$$\hat{p}(x;\hat{\theta}) = \arg \inf_{p(x;\theta)\in\Pi(\theta)} \inf_{\theta\in\Theta} I(p(x;\theta) \,||\, r(x;\theta)).$$

Thanks to the convex duality, the estimator $\hat{\theta}$ can equivalently be obtained as

$$\hat{\theta} = \arg \inf_{\theta\in\Theta} \inf_{\lambda\in\mathbb{R}^J} \log \sum_{i=1}^{m} r(x_i;\theta) \exp\left(-\sum_{j=1}^{J} \lambda_j u_j(x_i;\theta) \right).$$

The estimator is known as Maximum Maximum Entropy (MaxMaxEnt) estimator.

Empirical parametric Π-problem: Empirical MaxMaxEnt

The parametric Π-problem can be made more realistic, by assuming that a sample of size N is available to a modeler. The sample can be used to estimate the convex dual objective function by its sample analogue

$$\log \sum_{l=1}^{N} \exp \left(- \sum_{j=1}^{J} \lambda_j u_j(x_l; \theta) \right) .$$

The resulting method is known as Empirical Maximum Maximum Entropy (EMME) method, or Maximum Entropy Empirical Likelihood.

Summary

The Φ-problem is a statistical, bayesian problem of selection of a sampling-distribution from model set Φ, when a sample is observed. Method of selection should be consistent. Hence, it has to satisfy the BLLN. MNPL/EL and MAP satisfy the BLLN; extremization of other discrepancy measures or selection of posterior mean are, in general, inconsistent methods.

The Π-problem is a problem of selection of an empirical distribution from model set Π, when sampling distribution is known. Method of selection should be consistent. Hence, it has to satisfy the CLLN. MaxProb and REM satisfy the CLLN; extremization of other discrepancy measures (like Rényi-Tsallis entropy) is, in general, inconsistent.

Acknowledgement

Discussions and cooperation with George Judge (UC, Berkeley) and Robert Niven (UNSW, Canberra) are gratefully acknowledged.

Special thanks to Valérie Girardin.