Criterion selection for Boltzmann Jaynes Inverse Problem: concluding considerations

M. Grendar

Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, Bratislava, 841 04, Slovakia. KM FPV UMB, Banská Bystrica, Slovakia.

Abstract. *Boltzmann-Jaynes ill-posed inverse problem (BJIP) is usually regularized by means of extremization of a criterial function. When BJIP is defined by a moment consistency constraint (mcc) it is natural to place on the criterial function a requirement of complementarity of the extremization problem with Maximum Likelihood task. Previously, it was shown that Kullback-Leibler (KL) criterion satisfies the ML-complementarity requirement, while an Empirical Likelihood kind of criterion is not ML-complementary. Here, it is shown that KL is the only criterion which satisfies the ML-complementarity requirement. This result at the same time completes a non-axiomatic justification of Relative Entropy Maximization (REM/MaxEnt) method, since it shows that REM/MaxEnt is the only method which under mcc transforms potential linearly.*

Keywords: *Boltzmann Jaynes Inverse Problem, Relative Entropy Maximization*

1. Introduction

Boltzmann-Jaynes inverse problem (BJIP, [8]) arises wherever quadruple of information $\{\mathcal{X}, n, q, \Pi_n\}$ is available and it is necessary to select one (or more) types from the feasible set Π_n. In the quadruple, $\mathcal{X}$ is m-element support of random variable X, of which sequences of length n are assumed to induce types (i.e., empirical distributions, occurrence vectors) from set Π_n; q is the (supposed or true) source of sequences and hence also of types. BJIP can be met in many branches of science and engineering. For instance, the problem of 'recovering' a probability distribution from observations which were measured within known intervals falls into BJIP category.

In order to solve BJIP (i.e., to select from Π_n a type, given the information quadruple) some criterion (or objective function) is usually employed. This consequently turns the problem of solving BJIP into a problem of selecting the criterion, and justifying choice of the criterion. At [4], [5] a meta-criterion for judging 'reasonability' of the criterion was proposed for the case of the most common instance of BJIP at which the feasible set is defined by the moment consistency constraints (mcc). There, it was suggested that the criterion choice problem can be in this case solved by means of Maximum Likelihood (ML) complementarity meta-criterion.

Relative Entropy Maximization (REM/MaxEnt, cf. [9]) criterion/method is frequently used for solving BJIP. At [4], [5] it was shown that under mcc REM/MaxEnt is complementary with ML (in other words: REM/MaxEnt satisfies the ML-complementarity meta-criterion). Acceptance of REM/MaxEnt as 'the only proper criterion' for solving BJIP is not equivocal, however. There are several other criteria in use. An empirical-likelihood sort of criterion was at [5] shown to not satisfy the ML-complementarity meta-criterion.

The main objective of this paper is to show that if Π is defined by mcc then REM/MaxEnt is the only criterion which satisfies the ML-complementarity requirement.

The ML-complementarity approach to the criterion choice problem was formulated geometrically at [6] where it served also as a basis for a non-axiomatic approach to justification of REM/MaxEnt method in the case of the feasible set defined by linear moment consistency constraints. The justification arises from the fact that REM/MaxEnt transforms a potential linearly - hence, in the simplest possible way. An opened question remained, whether REM/MaxEnt is the only method with this property. This work shows that answer to the question is affirmative.

2. Criterion Choice Problem: ML-complementarity approach

First, moment consistency constraints should be recalled. Let there be a random sample of sufficiently large size n which supposedly or truly comes from q, and a given function (called potential) $u(X)$ of the random variable X. Based on them, an average a of the potential u is $a = \sum_{i=1}^{m} \nu_i^n u_i$ where ν_i^n is relative occurrence of i-th element of $\mathcal{X}$ at the sample (i.e., ν^n is type induced by the random sample). Moment consistency constraint equates together expected value of u and the average a, and this way it forms a feasible set Π of *probability mass functions* (pmf's) p; $\Pi = \{p : \sum_{i=1}^{m} p_i u_i = a\}$.

BJIP with Π defined by mcc (mcc-BJIP) can be solved by means of a criterion $C(p)$. Let $\mathcal{C}$ be a class of all criterial functions $C(p)$ such that 1) $C(p)$ is in order to solve mcc-BJIP extremized subject to mcc, and 2) First Order Conditions of the extremum lead to a family of pmf's $p(\lambda)$ parametrized by a lagrange multiplier λ of the constrained extremum problem. Let the value of the parameter λ which satisfies mcc be $\hat{\lambda}$.

Provided that we restrict ourselves in solving BJIP with mcc to criteria from the class $\mathcal{C}$, it is natural to require that a criterial function $C(p)$ satisfies a requirement of complementarity with a Maximum Likelihood task, cf. [4], [5], [6]. It is instructive to break up formulation of ML-complementarity meta-criterion into two parts: 1) a self-standing requirement for the ML task, 2) a requirement for mcc-BJIP:

1) ML task: Let there be a random sample of sufficiently large size n such that it induces type for which $\sum_{i=1}^{m} \nu_i^n u_i = a$. It is assumed that source of the sample lays in a parametric family $p(\lambda)$ which is parametrized by single parameter λ. Let λ_0 be the value of λ such that $f(\lambda_0) \in \mathcal{Q}$, where $\mathcal{Q} = \{q : E_q u(X) = a\}$. Let $\tilde{\lambda}$ be the sample-based ML estimate of λ. It seems reasonable to put on the form of the parametric family $f(\lambda)$ such a requirement that $p(\tilde{\lambda})$ should have value of the theoretical u-moment $E_{p(\tilde{\lambda})} u(X)$ equal to a; i.e., $p(\tilde{\lambda})$ should solve mcc. In other words, $\tilde{\lambda} \equiv \lambda_0$. This will be called "Requirement of internal consistency of ML task".

2) BJIP with mcc asks to select a probability mass function p from $\Pi = \{p : \sum_{i=1}^{m} p_i u_i = a\}$. A criterion $C(p) \in \mathcal{C}$ for selection of pmf should be such that it will under mcc choose such a family of pmf's $p(\lambda)$ which – when used at the ML task – satisfies the Requirement of internal consistency of ML task, for any random sample of sufficiently large size n which induces type ν^n from $R_n = \{\nu^n : \sum_{i=1}^{m} \nu_i^n u_i = a\}$. Note, that this in turn guarantees that $p(\hat{\lambda}) = p(\tilde{\lambda})$.

3. REM/MaxEnt as the only ML-complementary method

It remains to show that $C(p) = \pm \sum_{i=1}^{m} p_i \log(p_i/\gamma_i)$ is the only criterial function from $\mathcal{C}$ which satisfies the ML-complementarity meta-criterion. It will be done in two steps. First, it will be shown that $p(\lambda)$ should have the form of $c\gamma_i e^{-\lambda u_i}$ in order to satisfy the Requirement of internal consistency of ML task. At the next step, it will be shown that extremization of $C(p) = \pm \sum_{i=1}^{m} p_i \log(p_i/\gamma_i)$ under mcc leads to the above parametric family.

Step 1: Given a random sample/type r from R_n, and supposed parametric family $p(\lambda)$ from which the sample was drawn the ML task is to find ML estimator $\hat{\lambda}$ of λ as $\arg\max_\lambda \sum_{i=1}^m r_i \log p_i(\lambda)$. Since $p_i(\lambda)$ can be written as $p_i(\lambda) = k(\lambda) f(\lambda, x_i)$, where $k(\lambda) = 1/\sum_{i=1}^m f(\lambda, x_i)$ and $p'_\lambda = k'_\lambda f + k f'_\lambda$ the score equation (i.e., FOC) becomes

$$\sum_{i=1}^m r_i \frac{k'_\lambda}{k} = -\sum_{i=1}^m r_i \frac{f'_\lambda}{f}. \tag{1}$$

The score equation (1) should be identical with mcc:

$$\sum_{i=1}^m p_i u_i = \sum_{i=1}^m r_i u_i. \tag{2}$$

Since k, and k'_λ are constants, equating LHS of (1) and (2) leads to requirement

$$\frac{k'_\lambda}{k} = \sum_{i=1}^m p_i u_i. \tag{3}$$

Equating RHS of (1) and (2) leads to requirement

$$\frac{f'_\lambda}{f} = -u_i. \tag{4}$$

Requirements (3) and (4) are satisfied only by the exponential family $p_i(\lambda) = k(\lambda)\gamma_i e^{-\lambda u_i}$, where $\gamma_i \in \mathrm{R}$.

Step 2: It is necessary to find out a criterial function $C(p) \in \mathcal{C}$, extremization of which under mcc leads to the exponential family of pmf's

$$p_i(\lambda) = k(\lambda)\gamma_i e^{-\lambda u_i}. \tag{5}$$

Lagrangean $C(p) + \lambda(a - \sum p_i u_i) + \mu(1 - \sum p_i)$ is extremized wrt p_i for a p_i which satisfies

$$C'_{p_i} = \lambda u_i + \mu. \tag{6}$$

At the same time (5) implies that $-\log p_i = \lambda u_i - \log k(\lambda) - \log \gamma_i$. This, compared with (6) shows that 1) $\log k(\lambda) = -\mu$ and 2) $C'_{p_i} = -\log p_i + \log \gamma_i$. (The other possibility, namely that $\log k(\lambda) + \log \gamma_i = -\mu$ and $C'_{p_i} = -\log p_i$ would imply that $C(p) = -p_i \log p_i$ which in turn would lead through the extremization of the lagrangean to conclusion that $-\mu = 1 + \log k(\lambda)$, which implies that $\log \gamma_i = 1$ and leads to nowhere). The second requirement then implies that $C(p) = -p_i \log(p_i/\gamma_i) + p_i + d$.

It is straightforward to see that this $C(p)$ should be maximized ($-C(p)$ minimized), in order to attain the ML-complementarity. Since the feasible set is convex and the criterial function is concave (convex) the maximum (minimum) is unique.

4. Notes

1) A meta-criterion which leads to the same conclusion as the ML-complementarity meta-criterion, has already been considered in the literature. Based on an analogy with the directed orthogonality principle for linear projections in Hilbert space, Jones [10] advocated a directed orthogonality meta-criterion for

criterion selection in mcc-BJIP. Earlier, Campbell [1] demonstrated that under mcc-BJIP, REM/MaxEnt and a method based on Gauss's principle are equivalent.

2) Moment consistency constraint equates theoretical and empirical moment. This is reasonable if n - the sample size - is either sufficiently large or not known. If n is known and not 'sufficiently large', then one should consider BJIP with set Π_n in place of Π (then, clearly, REM/MaxEnt does not lead to the exponential family). However, it is (strangely) more common, to consider BJIP with Π defined by mcc even when n is finite and known. Then, MaxEnt/REM under mcc remains ML-complementary, regardless of n. Objective functions of ML task with exponential family and REM/MaxEnt task under mcc, however, attain a compatible relationship only for $n \to \infty$, cf. [7].

3) Though nothing restricts γ to be a pmf it seems natural to identify γ with q, since $p(\lambda)$ should be also a function of q, regardless of whether q is viewed as the true or supposed source.

Note that each of the views of q implies different interpretation of mcc-BJIP. If q is taken to be the true source of the sufficiently large random sample and the sample average of $u(X)$ is different than $E_q u(X)$ then it implies that a large deviation occurred. In mcc-BJIP, where the sample is not given to us, we are thus interested in guessing a structure of the deviation. On the other hand, if q is viewed as an a-priori source, then mcc-BJIP can be be viewed as a problem of 'induction'.

4) It should be stressed out, that REM/MaxEnt was shown to be the exclusive ML-complementary method, under the linear moment consistency constraints, only. The ML-complementarity meta-criterion is itself applicable for selecting among criteria from class $\mathcal{C}$, only. Thus, other approach should be taken in order to justify REM/MaxEnt under a broader setting. Indeed, there exist axiomatic (cf. [2], [11]) as well as a probabilistic justifications [3] of REM/MaxEnt which are not limited to mcc instance of BJIP.

5) The mcc defines feasible set Π which is convex. If $C(p) \in \mathcal{C}$ is a convex (concave) function of p then there exists to the problem of its minimization (maximization) under mcc a dual problem. It is worth noting that if the criterial function is Kullback-Leibler distance (relative entropy) then the complementary ML task is formally identical to the convex dual problem. For any other convex (concave) criterial function the respective ML task is not complementary (as it was shown here), yet its convex dual problem exists.

6) The above argument could be made also for continuous random variable X, at a price of involved technicalities. The argument directly generalizes to the case of several moment consistency constraints.

5. Acknowledgements

The work was supported by the grant VEGA 1/0264/03 from the Scientific Grant Agency of the Slovak Republic.

References

[1] Campbell, L. L., "Equivalence of Gauss's principle and Minimum discrimination information estimation of probabilities,", *Ann. Math. Stat.*, 41, pp. 1011-1015, 1970.

[2] Csiszár, I., "Why Least Squares and Maximum Entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, 19, pp. 2032-2056, 1991.

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

[4] Grendár, M., Jr. and Grendár, M., "MiniMax Entropy and Maximum Likelihood: Complementarity of Tasks, Identity of Solutions," in *Bayesian inference and Maximum Entropy methods in Science and Engineering*, A. Mohammad-Djafari (ed.), pp. 49–61, AIP Press, Melville, NY, 2001.

[5] Grendár, M., Jr. and Grendár, M., "ML-complementarity approach to Criterion Choice Problem," in *Measurement 2001*, I. Frollo, M. Tyšler and A. Plačková, (eds.), pp. 261-265, IMS SAS, 2001.

[6] Grendár, M., and Grendár, M., Jr., “Why Maximum Entropy? A non-axiomatich approach,” in *Bayesian inference and Maximum Entropy methods in Science and Engineering*, R. L. Fry (ed.), pp. 375-379, AIP, 2002. Also at arXiv:math-ph/0212005.

[7] Grendár, M., Jr. and Grendár, M., ”Chernoff’s bound forms,” in *Bayesian inference and Maximum Entropy methods in Science and Engineering*, Ch. Williams (ed.), AIP Press, Melville, NY, 2003.

[8] Grendár, M., Jr. and Grendár, M., ”Maximum Entropy method with non-linear moment constraints: challenges,” in *Bayesian inference and Maximum Entropy methods in Science and Engineering*, G. Erickson and Y. Zhai (eds.), pp. 97-109, AIP, Melville, 2004. Also at arXiv:physics/0308006.

[9] Jaynes, E. T., *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkratz (ed.), D. Reidel, Dordrecht, 1979.

[10] Jones, L. K., ”Approximation-theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the Maximum Entropy method to incorporate prior knowledge,” *SIAM J. Appl. Math.*, 49, pp. 650-661, 1989.

[11] Paris, J. and Vencovská, A., ”In defense of Maximum Entropy inference process,” *Internat. J. Approx. Reason.*, 17, pp. 77-103, 1997.