Empirical Maximum Entropy Methods

M. Grendar* and G. Judge†

*Department of Mathematics, FPV UMB, Tajovskeho 40, 974 01 Banska Bystrica, Slovakia;
Institute of Measurement Science, Bratislava; Institute of Mathematics and CS, Banska Bystrica.
Email: marian.grendar@savba.sk.
†207 Giannini Hall, University of California, Berkeley, CA, 94720.
Email: judge@are.berkeley.edu.

Abstract. A method, which we suggest to call the Empirical Maximum Entropy method, is implicitly present at Maximum Entropy Empirical Likelihood method, as its special, non-parametric case. From this vantage point the empirical approach to estimation is surveyed.

INTRODUCTION

Relative Entropy Maximization method (REM/MaxEnt) can be used at few contexts [7], for different purposes and its use there can be justified by various arguments. In this note we consider REM within the framework of Boltzmann Jaynes Inverse Problem (the α-problem, for short). The problem is framed by information-quadruple $\{\mathcal{X}, r, n, \Pi\}$ where $\mathcal{X}$ is support of random variable X with probability mass function (pmf) r, and Π is a set of pmf's into which empirical measure (hereafter called also type, or n-type) – induced by an unavailable random sample from r of size n – is known to belong. The objective is to select a type (one or more) from Π when the information-quadruple and nothing else is available. If Π contains more than one type, the α-problem becomes under-determined and in this sense ill-posed. Application of REM to the α-problem can be justified on probabilistic grounds, via Conditional Limit Theorem (CoLT) [21], [20], [7], Gibbs Conditioning Principle [6], [8] and/or as asymptotic instance of Maximum Probability method [9].

The α-problem is an idealization, since at least r and Π are usually not known. The feasible set Π should be constructed some way, and the 'true' data-sampling pmf r should be 'guessed', hence r is in practice replaced by some 'reasonable' q. The most commonly considered feasible set Π is the one defined by moment consistency constraints (mcc) $\Pi \triangleq \{p : \sum p_i u_j(x_i) = a_j, j = 1, 2, \ldots, J\}$, where u's are real-valued functions and the vector $a \in R^J$ contains the sample-based values of the J-tuple of the left-hand-side u-moments. Researcher should specify the functions u, called potentials, which are thought to be a characteristic of the studied phenomenon. Choice of the so-called source q usually arises from *a priori* considerations.

EMPIRICAL MAXIMUM ENTROPY METHODS

In this work we extend the common approaches to both construction of Π and selection of q. In particular, we assume that there is available a random sample of size N from the 'true' data-sampling distribution r, and the guess/estimate q of r can thus be based on it. This way the original α-problem turns into *empirical* α-problem. In the case of discrete random variable the estimate q can be constructed directly. The challenge of continuous case can be handled in two ways (see the next section).

The standard mcc construction of the feasible set Π can become more versatile, if instead of the potential function $u(x)$ its parametric extension $u(x, \theta)$ is considered. This way the α-problem with mcc turns into parametric mcc α-problem.

We gradually blend together these two extensions in order to get the empirical parametric mcc α-problem, and note that it has to be solved by Empirical Maximum Maximum Entropy method (EMME), known from the econometric literature. The simpler problems which it encapsulates are at best implicit in the literature.

Empirical MaxEnt

The simplest of the problems, the parametric α-problem can be illustrated by the following example.

Example. *Let $\mathcal{X} = \{1,\ 2,\ 3,\ 4\}$, $\Pi = \{p : \sum p_i x_i = a\}$ and let $n = 10^9$ so that the feasible set of types can be effectively considered to be set of pmf's. The value of the u-moment based on random sample of size n which is not available to us was found to be $a = 3.1$. The data-sampling pmf r is not known to us, either. However, instead of r, let there be a random sample of size $N = 40$ from r, which induced empirical pmf $\nu^N = [5,\ 14,\ 12,\ 9]/40$. The objective is to select an n-type (effectively a pmf) from Π, given the available information. We thus face the empirical α-problem with feasible set defined by the moment-consistency constraint.*

It is reasonable to estimate the data-sampling pmf r by the empirical pmf ν^N. Conditional Limit Theorem then implies that the empirical α-problem should be solved by selecting the information projection of ν^N on Π. We call the associated method Empirical Maximum Entropy (EME). As $N \to \infty$, the information projection converges (a.s.) to information projection of r on Π.

Continuous case

Assume now that the underlying random variable X is continuous, with probability density function $r(X)$, unknown to us. The continuous-case analogue of the Example has the following form: $n = 10^9$, $\Pi = \{p : \sum_{l=1}^{n} p_l x_l = a\}$ where $a = 3.1$. Again, let there be a random sample of size N drawn from $r(X)$, which could provide an estimate of $r(X)$. The question is, how? One possibility is to use kernel estimator. This would however introduce a new source of 'uncertainty' into the problem.

Let us recall that in the continuous case CoLT dictates to solve the empirical α-problem by selecting the information projection $\hat{p} = \arg\inf_{p\in\Pi} I(p||q)$, where now $I(p||q) \triangleq \int p\log(p/q)$, and q is the data-based pdf-estimate of r. There are two ways considered in the literature how to obtain the information projection using directly the sample data, and thus avoiding use of an intermediate construct like the kernel smoother.

Let us begin with the more difficult yet more common one, which we call empirical estimation 'trick'. The trick [18] lays in forcing the observed random sample (of size N) to become the support $\mathcal{S}$ of a random variable S with the uniform distribution u. This is done by assuming that no two observed data-points are identical. (Alternatively, the trick could be explicated as forming a Dirac-type estimator of the continuous data-sampling distribution r.) The feasible set Π of pdf's thus turns into the set $\Pi_S \triangleq \{p_S(\cdot) : \sum_{l=1}^{N} p_S(s_l)u_j(s_l) = a_j, 1 \le j \le J\}$ of pmf's on the support $\mathcal{S}$. This way the continuous setting turned into the discrete-case empirical α-problem, and the discrete-case form of CoLT can be used to justify EME as the correct method of its solution. The method selects $\hat{p} = \arg\inf_{p\in\Pi_S} I(p||u)$. Note that the method if used in the discrete case would collapse into the discrete-case EME.

The other approach was used at [14]. It is based on the observation that the convex dual problem to $\hat{p} = \arg\inf_{p\in\Pi} I(p||r)$ leads to $\hat{p}(x;\hat{\lambda}) \propto q(x)\exp(\sum \hat{\lambda}_j u_j(x))$ where

$$\hat{\lambda} = \arg\sup_{\lambda\in R^J} \left\{ \sum \lambda_j a_j - \log \int r(x)\exp\left(-\sum_{j=1}^{J} \lambda_j u_j(x) \right) \right\}. \tag{1}$$

Since instead of the 'true' data-sampling distribution $r(X)$ we have only the N-sample, it is natural to replace (1) by its empirical analogue

$$\hat{\lambda} = \arg\sup_{\lambda\in R^J} \left\{ \sum \lambda_j a_j - \log \sum_{l=1}^{N} \exp\left(-\sum_{j=1}^{J} \lambda_j u_j(x_l) \right) \right\}. \tag{2}$$

The two approaches lead to the same result.

MaxMaxEnt

Scope of the standard moment-consistency α-problem can be expanded by replacing the constraints by parametric moment consistency constraints $\Pi(\theta) \triangleq \{p(\cdot,\theta) : \sum p(x_i,\theta)u_j(x_i,\theta) = 0, j = 1,2,\ldots,J\}$ where $\theta \in \Theta \subseteq R^k$. In this context it is worth envisioning the data-sampling distribution $r(X)$ as a member of parametric family of distribution $f(x,\theta)$. The functions u are in this context commonly known as estimating functions and the moment-consistency constraints are called unbiased estimating equations (of θ), for obvious reasons. The parametric α-problem is framed by the information-pentad $\{\mathcal{X}, r, n, \Pi(\theta), \Theta\}$, and the objective is now to select parametric type from $\Pi(\theta)$ when nothing else except of the pentad is available. It is thus necessary to select both p and θ. If selecting θ is of greater concern, the problem can be viewed as a problem of estimation of the 'true' value of the parameter θ of the data-sampling distribution $f(x,\theta)$.

An example could help to fix ideas.

Example. *Let $\mathcal{X} = [1 \;\; 2 \;\; 3 \;\; 4]$. Assume that it is known that $r(x)$ belongs to a parametric family $f(x,\theta)$, such that $u(x) = x$ is unbiased estimating function of the scalar parameter. Since $r(x)$ is not known to us, we make a guess $q = [0.1 \;\; 0.6 \;\; 0.2 \;\; 0.1]$ of it, based on a prior information. Data come in the form of information that a sample of size $n = 10^9$ (so that effectively we can replace types by pmf's) led to the sample value of the u-moment in the range $[2.3, 3.7]$. Thus, $\Pi(\theta) \triangleq \{p(\theta) : \sum_{i=1}^{4} p_i(\theta)(x_i - \theta) = 0\}$, where $\theta \in \Theta = [2.3, 3.7]$. Given the information-pentad $\{\mathcal{X}, r, n, \Pi(\theta), \Theta\}$, the objective is to select a pmf from $\Pi(\theta)$.*

CoLT implies that the parametric mcc α-problem should be solved by selecting $\hat{p}(\theta) = \arg\inf_{p \in \Pi(\theta)} I(p(\theta)||q)$, with $\theta = \hat{\theta}$, where $\hat{\theta} = \arg\inf_{\theta \in \Theta} I(\hat{p}(\theta)||q)$. Since the solution can be equivalently obtained as a double maximization of the relative entropy (over p and θ) it is reasonable to call the associated method Maximum Maximum Entropy (MaxMaxEnt).

The continuous-case version of the problem can be handled in either of the two ways, described in the previous section.

Empirical MaxMaxEnt

The parametric mcc α-problem can be extended into the more realistic empirical version, if it is assumed that in addition of the 'aggregated data' there is also a random sample of size N from the data-sampling distribution. The problem is made more flexible, by replacing the assumption that Θ is given as an aggregate based on sample of size n, by assumption that Θ is simply specified some way.

The empirical parametric mcc α-problem contains all the ingredients which are considered accessible in the current econometric research: $\mathcal{X}$, the sample (of size N), set of estimating equations and Θ.

If the above information and nothing else is supplied, CoLT dictates to solve the problem of selecting $p(\theta)$ by choosing $\hat{p}(\theta) = \arg\inf_{p \in \Pi(\theta)} I(p(\theta)||\nu^N)$, with $\theta = \hat{\theta}$, where $\hat{\theta} = \arg\inf_{\theta \in \Theta} I(\hat{p}(\theta)||\nu^N)$ and ν^N is the empirical measure induced by the sample. The associated method is known either as Maximum Entropy Empirical Likelihood [16] or Exponentially Tilted estimator [1], [2], [12], [14]; we call it Empirical Maximum Maximum Entropy (EMME) method. Its continuous-case variant can be constructed either via the empirical estimation trick or by the approach of [14]. In regards of the approach based on the empirical estimation 'trick' is is worth noting that the convex dual problem to $\hat{p} = \arg\inf_{\theta \in \Theta} \inf_{p \in \Pi(\theta)} I(p||u)$ can be written as:

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \sup_{\lambda} \frac{1}{n} \sum_{l=1}^{n} \log \hat{p}_S(s_l; \lambda, \theta),$$

where

$$\hat{p}_S(\cdot) = k(\lambda, \theta) \exp(-\sum \lambda_j u_j(s_l, \theta)),$$

and $k(\lambda, \theta)$ is the normalizing constant. The dual formulation shows that EMME is equivalent to a MiniMax Likelihood method[1], which utilizes likelihood function based on the exponential family and the data-induced type u.

Finally, CoLT implies that the available information cannot be processed by a method other than EMME. Thus, Empirical Likelihood or Generalized Method of Moments in this case violate the probabilistic theorem; for further discussion see [11].

SUMMARY

Empirical Estimation approach [17], [18], [19] can be combined with relative entropy as the regularizing criterion [1], [12], [14], [4], [15], see also [18] and [16]. The approach is used to process information of certain form in order to obtain estimator of unknown values of parameters of interest, so that consequently inferences can be made, cf. [18], [16]. The information comprises support $\mathcal{X}$, random sample of size N, set of estimating equations and parametric space Θ. As it was recognized at [15], Conditional Limit Theorem and Gibbs Conditioning Principle apply to the setting, and show that the information should be processed by means of the relative entropy maximization. The resulting method is known under various names; here Empirical Maximum Maximum Entropy (EMME) method. The setting and hence also associated EMME contain simpler settings and methods, which could be of some independent interest. They were discussed in this note.

ACKNOWLEDGEMENTS

M.G. gratefully acknowledges financial support of his participation at the MaxEnt06 workshop from the Jaynes Foundation. Supported (M.G.) by VEGA grant 1/3016/06.

REFERENCES

1. Back, K., Brown, D. (1990). Estimating distributions from moment restrictions. Working paper, Graduate School of Business, Indiana University.
2. Baggerly, K. A. (1998). Empirical Likelihood as a goodness-of-fit measure. *Biometrika*. 85/3:535-547.
3. Baggerly, K. A. (1999). Studentized Empirical Likelihood and Maximum Entropy. Technical report, Rice University, Dept. of Statistics.
4. Corcoran, S. A. (2000). Empirical exponential family likelihood using several moment conditions. *Stat. Sinica*. 10:545-557.
5. Cover, T., Thomas, J. (1991). *Elements of Information Theory*. New York:Wiley.
6. Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* 12:768-793.
7. Csiszár I., MaxEnt, mathematics and information theory, *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver (eds.), pp. 35-50, Kluwer, 1996.

---

[1] Confront it with the MiniMax Entropy method [10].

8. Dembo, A., Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. New York:Springer-Verlag.
9. Grendár M. Jr and Grendár, M., What is the question that MaxEnt answers: a probabilistic interpretation, in *Bayesian inference and Maximum Entropy methods in Science and Engineering*, A. Mohammad-Djafari (ed.), AIP, Melville (NY), pp. 83-94, 2001.
10. 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.), AIP, Melville (NY), pp. 49-61, 2001.
11. Grendár, M., Judge G. (2006) Large Deviations Theory and Empirical Estimator choice, preprint ARE Berkeley, 2006. http://repositories.cdlib.org/are_ucb/1012.
12. Imbens, G. W., Spady, R. H.,Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*. 66/2:333-357.
13. Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Phys. Rev.* 106:620 and 108:171.
14. Kitamura, Y., Stutzer, M. (1997). An information-theoretic alternative to Generalized Method of Moments estimation. *Econometrica*. 65:861-874.
15. Kitamura, Y., Stutzer, M. (2002). Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics*. 107:159-174.
16. Mittelhammer, R. C., Judge, G. G., Miller D. J. (2000). *Econometric Foundations*. Cambridge:CUP.
17. Owen, A. B. (1991). Empirical Likelihood for linear models. *Ann. Statist.* 19:1725-1747.
18. Owen, A. B. (2001). *Empirical Likelihood*. New York:Chapman-Hall/CRC.
19. Qin, J., Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist.* 22:300-325.
20. van Campenhout J. M. and Cover T. M., Maximum entropy and conditional probability, *IEEE IT*, 27, pp. 483-489, 1981.
21. Vasicek O. A., A conditional law of large numbers, *Ann. Probab.*, 8, pp. 142-147, 1980.