Modeling the Learning from Repeated Samples: a Generalized Cross Entropy Approach

Rosa Bernardini Papalia[1]

[1] Dipartimento di Scienze Statistiche, Università di Perugia, V. A. Pascoli, 06122 Perugia

Summary

In this study we illustrate a Maximum Entropy (ME) methodology for modeling incomplete information and learning from repeated samples. The basis for this method has its roots in information theory and builds on the classical maximum entropy work of Janes (1957). We illustrate the use of this approach, describe how to impose restrictions on the estimator, and how to examine the sensitivity of ME estimates to the parameter and error bounds. Our objective is to show how empirical measures of the value of information for microeconomic models can be estimated in the maximum entropy view.

Keywords: Generalized Maximum Entropy, Generalized Cross Entropy, Repeated Samples, Microeconometric models.

1 Introduction

In this study we illustrate a Maximum Entropy (ME) methodology for modelling incomplete information and learning from repeated samples. The basis for this method has its roots in information theory (Shannon, 1948) and builds on the classical maximum entropy work of Janes (1957). We illustrate the use of this approach, describe how to impose restrictions on the estimator, and how to examine the sensitivity of ME estimates to the parameter and error bounds. Generalized Maximum Entropy (GME) and Generalized Cross Entropy (GCE) estimation frameworks are presented.

The GME/GCE formulations are designed to introduce sample information in either a data or moment form, and they permit to make use of all available information. Specifically, the GCE approach proceeds by minimizing the entropy between a prior estimate and the reconstructed probability. If the generalized cross entropy measure is greater than zero we have gained information on the prior and thus learning has occurred. In the presence of repeated samples, cross entropy acts as a shrinkage rule so that the reconstructed probability approaches the true probability as the sample size approaches infinity (Golan et al., 1996). As would be expected if the correct prior information is available and it is employed within the estimation process this improves the accuracy of the estimation. Within this framework, minimal distributional assumptions are necessary and a dual loss function is used to take into account both the estimation precision and prediction objectives. However, the GME/GCE solutions behave like other shrinkage estimators. The parameter estimates are shrunk towards the prior mean, which is based on non sample information and thus as we increase the degree of shrinkage towards the prior mean we need to make sure that the prior mean is based on good nonsample information. Nevertheless, incorrect prior information does not significantly impact upon the accuracy of the estimation. The reason is because to achieve an interior solution to the problem the constraints must to be satisfied, but as the entropy method needs to satisfy the sample information any estimates will not stray too far.

The variance of the GME/GCE is less than the variance of sample-based rules like Least Squares or Maximum Likelihood (Kapur and Kesavan, 1992; Owen, 1991; Qin and Lawless, 1994), but the use of prior information introduces bias. Nevertheless, this bias is typically offset by variance reductions and the resulting mean squared error of the estimator is smaller than sample-based mean squared error.

The outline of the paper is as follows: the next section briefly describes the basic concepts of the maximum entropy estimation procedures. Section 3 introduces the both the GME and GCE estimation approache and also the framework for

modeling the learning from repeated samples. Finally, section 4 provides concluding remarks.

2 The Maximum Entropy principles

The principle of maximum entropy is based on: *(i)* the measure of the lack of information relative to a specific frame of reference; *(ii)* the measure of uncertainty, relative to the most informative probability distribution. The basic idea is to utilize the information available in an efficient way, even if partial and/or incomplete. Maximization of uncertainty relative to what is unknown is required and the most "uncertain" and "less probable" probability distribution is selected among those that are compatible with what is known and introduced in the form of restrictions in the data. The probability distribution that maximizes the entropy is the distribution that produces greater information among those that are coherent with the basic knowledge of the phenomenon under study.

A crucial assumption regarding the formulation of the maximum entropy problem concerns the parameters of the model obtained as the average of the estimated probability distributions. Specification of the density function of the population and knowledge of the data generating mechanism are not required.

Based only on knowledge of some moments of the probability functions relative to *(i)* model parameters, *(ii)* errors, *(iii)* unknown explanatory variables, the principle of maximum entropy is used to select a single distribution of probability in such a way that the information observed is satisfied. The estimates are produced by the solution of a constrained maximization problem that requires the use of numerical methods to derive analytical solutions.

The solution which is derived by this formulation agrees with the known information but it expresses maximal uncertanty in relation to all other things. If some non-sample information about the unknown probabilities is available, this can be expressed in terms of a prior probability distribution and as a consequence the accuracy of the estimates is improved. Following Kullback (1959) and using the principle of cross-entropy it is possible to incorporate prior beliefs about parameters into the estimation process.

To measure both the information relative to a system and the importance of the contribution of each individual observation and every single restriction introduced into the formulation of the ME problem, normalized entropy, $S(p)$, is used. $S(p)=(-\sum_s p_s \ln p_s)(/\ln n)$, with $S(p)\in [0,1]$, where p is the probability distribution of interest, $S(p)=0$ and $S(p)=1$ reflect: absence of uncertainty and complete uncertainty, respectively. Therefore, it is possible to establish if additional information or even restrictions in the data expressed in the form of restrictions produce a reduction of uncertainty and consequently a reduction in the basic uncertainty relative to the phenomenon.

As a relative measure of uncertainty, normalized entropy can be utilized to compare alternative formulations in order to choose the system of restrictions that

best contributes to the reduction of the uncertainty relative to the phenomenon under study.

3 The basic model

Our first objective is to employ the Generalized Maximum Entropy methodology to estimate multiple-equation statistical models.

The problem we present involves noisy unobserved data related to the units which are not selected and the objective is to estimate the unknown parameters of the statistical model as well as the unknown values of the unobserved variables (Smith, 1983; Copas and Li, 1997).

Indicating with Z* the vector of latent variables $\{z^*_i\}$, with $i \in C_0$ where C_0 represents the set of units that were not selected and with Y_1 the vector of observable variables $\{y_i\}$ corresponding to the set of selected units in sample C_1, the vector YZ^* $[(n_0+n_1)\times 1]$ is defined as:

$$YZ^*_i = \begin{cases} y_i & i \in C_1 \\ z^*_i & i \in C_0 \end{cases} \qquad (1)$$

where n_0 and n_1 represent the sample size of the two subsets respectively of non-selected and selected units, C_0 and C_1 in the sample.

The GME "reparameterization" of the model expressed in matrix terms is given by:

$$\begin{bmatrix} Y_1 \\ Z^* \end{bmatrix} = \begin{bmatrix} X_1 & \\ & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad (2)$$

$$\begin{bmatrix} Y_1 \\ Z^* \end{bmatrix} = \begin{bmatrix} X_1 & \\ & X_2 \end{bmatrix} \begin{bmatrix} Z^1 p^1 \\ Z^2 p^2 \end{bmatrix} + \begin{bmatrix} V^1 w^1 \\ V^2 w^2 \end{bmatrix} \qquad (3)$$

The hypotheses relative to the parameters, the terms of error and the latent variables present in the model are: $\beta_{1k} = \Sigma_m Z^1_{km} p^1_{km}$, $\beta_{2k} = \Sigma_m Z^2_{km} p^2_{km}$, $(k=1,...,K)$, $\varepsilon_{1i} = \Sigma_j V^1_j w^1_{ij}$, $(i \in C_1)$, $\varepsilon_{2i} = \Sigma_j V^2_j w^2_{ij}$, $(i \in C_0)$, $z^*_i = \Sigma_h C_i r_{ih}$, $(i \in C_0)$. In correspondence to every parameter, every error and every unobservable dependent variable a limited discrete random variable with finite support is

introduced. We then proceed to estimate the probability distributions associated with each discrete variable introduced into the formulation in the hypothesis that all of the unknown parameters are centered on the support space.
In our case the probability distributions, given the support spaces we have established, are: p^1_k , p^2_k , w^1_i , $i \in C_1$, w^2_i , ($i \in C_0$, r_i , $i \in C_0$), where:
$Z^1_k = [Z^1_{k1}, Z^1_{k2}, ..., Z^1_{kM1}]'$, $Z^2_k = [Z^2_{k1}, Z^2_{k2}, ..., Z^2_{kM2}]'$
$V^1_i = [V^1_{i1}, V^1_{i2}, ..., V^1_{iJ1}]'$, $V^2_i = [V^2_{i1}, V^2_{i2}, ..., V^2_{iJ2}]'$,
$C_i = [C_{i1}, C_{i2}, ..., C_{iH}]'$, with M_1, M_2, J_1, J_2, H all greater than two.
The definition of support spaces Z^1, Z^2 for the k parameters of the two equations is carried out considering possible "a priori" knowledge of the sign and the possible value of every parameter and considering uniform intervals symmetric around zero.
In a similar way, we proceed to the definition of the support spaces of every single error of the two equations, choosing the lower and upper bounds of the support space according to the rule of the "three sigma" of Pukelsheim (1994). In this case we assume that the errors are extracted from a uniform distribution with mean zero and variance equal to $(y_{max} - y_{min})^2/12$. As the values of the objective variable for the sample's non-selected units are unknown, we assume the values of Z* are distributed uniformly in their support of the definition and, therefore, based on knowledge of both the percentage of the non-selected units and the support of variable Y_1, we calculate the variance of the uniform distribution.

Alternately, an "Empirical Bayes" type procedure can be used (Efron-Morris, 1973; Casella, 1985; Judge et al, 1990). Choosing definition supports of the discrete random variables present in the model is equivalent to imposing "a priori" restrictions on the parameter estimates, possibly in a way which is consistent with the economic theories below.

The choice of the support spaces represents a way to impose prior restrictions on the parameter estimates but inequalities can also be specified directly in the constraint set. As an alternative, prior information can be also imposed using the cross entropy approach (Kullback, 1959; Levine, 1980).
Note that in our formulation we have reparametrised z^* and we are interested in estimating the probability distribution for the unobserved z^*. We have specified a support for z^* and estimated the corresponding probabilities. The formulation we adopted means that in maximizing entropy, z^* appears in the constraints of the optimization problem and in the criterion function. It is also possible to treat z^* differently and we could obtain only point estimates of those unknown instead of estimating the probability distribution for the unobserved z^*.

Following the multistage GME formulation presented in Bernardini-Filippucci (2000) in the first stage the ME objective function is given by:

$$\max_{p^{i'},w_1,w_2,r} H(p^{i'},w_1,w_2,r) = -\sum_k \sum_m p^{i'}{}_{km}' \ln p^{i'}{}_{km} - \sum_i \sum_j w_{1ij}' \ln w_{1ij} - \\ - \sum_i \sum_j w_{2ij}' \ln w_{2ij} - \sum_i \sum_s r_{is}' \ln r_{is} \tag{3}$$

subject to:
(i) data consistency conditions:

$$\begin{aligned} &\sum_k \sum_m \left(X_{1ik}\, p^{i'}{}_{km}\, z_m\right) + \sum_j \left(w_{1_{ij}} v_{1_i}\right) = E = Y \\ &\sum_k \sum_m \left(X_{2ik}\, p^{i'}{}_{km}\, z_m\right) + \sum_j \left(w_{2_{ij}} v_{2_i}\right) = E \le Z^* \end{aligned} \quad ; \tag{4}$$

(ii) adding-up constraints:

$$\begin{aligned} &\sum_m p^{i'}{}_{km} = 1, \quad k = 1,..., K \qquad \sum_j w_{1_{ij}} = 1, \quad i = 1,..., n_1 \\ &\sum_j w_{2_{ij}} = 1, \quad i = 1,..., n_2 \qquad \sum_s r_{is} = 1, \quad i = 1,..., n_2 . \end{aligned} \tag{5}$$

The solution to the system of equations related to the fist-order condition produce the point estimates:

$$\begin{aligned} &\beta^{i'}{}_k = \sum_{m=1}^{M} p^{i'}{}_{km}\, z^{i'}{}_m \qquad i' = 1,2 \quad k = 1,.., K \\ &\varepsilon^{i'}{}_i = \sum_{j=1}^{J} w_{ij}^{i'} v_i^{i'} \qquad i = 1,.., N \quad i' = 1,2. \end{aligned} \tag{6}$$

These first step estimates are used in the second stage of the proposed GME formulation to replace the unobserved Z* in the consistency relation (4), which become: $X_2 Z\hat{p} + V_2 \hat{w}_2 = \hat{Z}^* \ge X_2 Zp + V_2 w_2$.

The GME estimator relative to our formulation, is consistent and asymptotically normal (Bernardini, 2002); such asymptotic properties are proven on the basis of the results presented by Golan et al. (1996, 1997) and Mittelhammer and Cardell (1996), assuming the following conditions:

1. the error's supports for each equation are symmetric around zero;

2. the definition supports relative to the vectors of the parameters β_1 e β_2 of the two equations have respectively as lower and upper limits the values (Z^1_{k1} e Z^1_{kM1}) for β_1 and (Z^2_{k1} e Z^2_{kM2}) for β_2;

3. the errors are independently and identically distributed; such hypotheses do not imply hypotheses of uncorrelation of the errors between the two equations;

4. *plim* (1\N) X'X exists and is not singular, where X is a block diagonal matrix consisting of X_1 and X_2, $N=n_1+n_2$.

To estimate the variances of every parameter, a resampling inference approach such as the *jackknife* (Hinkley, 1986; Wu, 1986) or the *bootstrap* (Efron, 1979) can be used.

The estimated standard errors need to be treated with caution for two reasons. First, as *(X'X)* is used in the calculation of the standard errors, the underlying collinearity in the design matrix will increase the associated variability of the estimated coefficients. The variance of the multipliers and residuals will be increased by the existence of the collinearity, such that each of the terms in the GME/GCE variance/covariance matrix increase with the degree of collinearity. However, this effect can be mitigated by specifying tight supports on the parameters.

Transforming GME into a GCE formulation when both the reference distributions for errors and latent variables are uniform yields the following objective function:

$$\min_{p^{i'},w_1,w_2,r;q^{i'}} H(p^{i'},w_1,w_2,r;q^{i'}) = \sum_k \sum_m p^{i'}{}_{km}' \ln p^{i'}{}_{km} - \sum_k \sum_m p^{i'}{}_{km}' \ln q^{i'}{}_{km} + \\ + \sum_i \sum_j w_{ij}' \ln w_{ij} + \sum_i \sum_s r_{is}' \ln r_{is} \qquad (7)$$

The major difference between Eq. (3) and (7) is that the unknown probability distributions $p^{i'}$ is now subject to prior information $q^{i'}$.
Now, suppose that the previous estimates $p^{i'}{}_{t-1}$ is the prior distribution $q^{i'}$, the resulting GCE problem is written as:

$$\min_{p_t^{i'},w_1,w_2,r;q^{i'}} H(p_t^{i'},w_1,w_2,r;p_{t-1}^{i'}) = \sum_k \sum_m p_t^{i'}{}_{km}' \ln p_t^{i'}{}_{km} - \sum_k \sum_m p_t^{i'}{}_{km}' \ln p_{t-1}^{i'} + \\ + \sum_i \sum_j w_{ij}' \ln w_{ij} + \sum_i \sum_s r_{is}' \ln r_{is} \quad (8)$$

With this model it is then possible empirically model the learning that occur from repeated samples. These results can be incorporated into the model defined in Eq. (3)-(5).

4 Concluding Remarks

In this work, a system of restrictions suitable to represent the uncertainty relative to the partial-incomplete economic data generation process is formulated. Alternative formulations of GME and GCE are described. An empirical method for deriving the value of information and learning is also presented and a GCE formulation for repeated samples is proposed.

As regards traditional estimation techniques, the formulation of the constrained maximization problem, in the maximum entropy view, does not require the use of restrictive parametric assumptions on the model; hypotheses regarding the form of the distribution of the objective variables are not formulated. Restrictions expressed in terms of inequality can be introduced and it is possible to calibrate the precision in the estimation. Moreover, good results are produced in the case of small-sized samples, in the presence of high numbers of explanatory variables, which are also highly correlated.

By means of a relative measure of uncertainty, Normalized Entropy, different cases and scenarios can be compared, verifying the informative contribution of every restriction introduced on the basis of the information available and the "knowledge" relative to the formation process of the unknown data.

The extra-sample information is used to define both the range of variation of the parameter values and the restrictions to be introduced into the optimization phase of the estimation procedure.

However, we cannot fail to mention several critical aspects connected with the maximum entropy estimation procedure such as: *(i)* the need to employ discrete probability distributions; *(ii)* the choice of the support space for the error terms and for the parameters of the model; *(iii)* hypotheses relative to the weight of the measures of entropy that make up the objective function.

References

Bernardini Papalia R., and C. Filippucci (2000), Inference from non-random samples: a maximum entropy approach, *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, New York, 1686-1691.

Bernardini Papalia R., (2002), Analyzing partial-incomplete economic data generation processes using maximum entropy, *Proceedings of the International Conference on Improving Surveys*, Copenhagen.

Casella G., (1985), An introduction to empirical Bayes data analysis, "*The American Statistician"*, 39, 83-87.

Copas J. B., and H. G. Li (1997), Inference for Non-random Samples, " *Journal of the Royal Statistical Society B*", 59, 55-95.

Csiszar I. (1991), Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, "*The Annals of Statistics*", 19, 2032-2066.

Efron B., and C. Morris (1973), Stei's estimation rule and its competitors- an empirical Bayes approach, "*Journal of the American Statistical Association*", 68**,** 117-130.

Efron B., (1979), Bootstrapping Methods: Another Look at the Jackknife, *"Annals of Statistics"*, 7, 1-26.

Golan A., Judge G., and D. Miller (1996), *Maximum entropy econometrics: robust estimation with limited data*, Wiley.

Golan A., Judge G. and D. Miller (1997), The Maximum Entropy Approach to Estimation and Inference: An Overview, in T. B. Fomby and R. C. Hill (eds.*), Applying Maximum Entropy to Econometric Problems*, Jai Press Ltd. England, London, 3-24.

Golan A., Judge G., and D. Miller (1997), Information recovery in simultaneous equation models, in Ullah A., Giles D.(editors), *Handbook of Applied Economic Statistics*. Marcel Dekker, New York.

Golan A., Judge G., and J. Perloff (1996), Recovering Information from Multinomial Response Data, "*Journal of the American Statistical Association*", 9**,** 841-583.

Hinkley D. (1986), Discussion of 'Jackknife, Bootstrap and other Resampling Methods in Regression Analysis', by C. F. J. Wu, "*Annals of Statistics"*, 14, 1312-1316.

Janes E. T. (1957), Information theory and statistical mechanics, "*Physics review*", 106, 620-630.

Judge G. G., Hill R. C., and M. E. Bock (1990), An adaptive empirical Bayes estimator of the multivariate normal mean under quadratic loss, *"Journal of Econometrics"*, 44, 189-213.

Kapur J.N., and H.K. Kesavan (1992), *Entropy optimization principles with applications*, Academic Press, INC.

Kullback J. (1959), *Information theory and statistics*, Wiley, New York, NY.

Levine R.D. (1980), "An information theoretical approach to inversion problems", *Journal of Physics* A, 13,91-108.

Mittelhammer R., and S. Cardell (1996), On the Consistency and Asymptotic Normality of the Data Constrained GME Estimator of the GML, Working paper, Washington State University.

Owen A. B., (1991), Empirical Likelihood for linear models, *"The Annals of Statistics"*, 19, 1725-1745.

Pukelsheim F. (1994), The Three Sigma Rule, "*American Statistician*", 48, 88-91.

Qin J., and J. Lawless (1994), Empirical Likelihood and General Estimating Equations, *"The Annals of Statistics"*, 22, 300-325.

Shannon C.E. (1948), A mathematical theory of communication., "*Bell System Technical Journal*", 27, 379-423.

Smith T.M.F. (1983), On the validity of inferences from non-random sample, "*Journal of the Royal Statistical Society A*", 146, 394-403.

Skilling J. (1989), The Axioms of Maximum Entropy, in J. Skilling (editor), *Maximum Entropy and Bayesian Methods in Science and Engineering*, Kluwer Academic, Dordrecht, 173-187.

Wu C. F. J. (1986), Jackknife, Bootstrap and other Resampling Methods in Regression Analysis', *"Annals of Statistics"*, 14, 1261-1350.