Physica A 453 (2016) 162–172

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Generalized Cross Entropy Method for estimating joint distribution from incomplete information

Hai-Yan Xu [a], Shyh-Hao Kuo [a], Guoqi Li [a,b], Erika Fille T. Legara [a], Daxuan Zhao [a,c], Christopher P. Monterola [a,*]

[a] Complex System Group, Department of Computing Science, Institute of High Performance Computing, 138632, Singapore
[b] Department of Precision Instrument, Tsinghua University, Beijing, 100084, China
[c] School of Businesses, Renmin University of China, Beijing, 100872, China

HIGHLIGHTS

- A new and novel algorithm named Generalized Cross Entropy Model (GCEM) is proposed.
- GCEM estimates full joint distribution from marginal even with incomplete information.
- Existing maximum entropy procedures are shown to be just special cases of GCEM.
- Accuracy of GCEM is established and illustrated using actual empirical data.
- Article provides guide on how GCEM could be applied to diverse fields/areas.

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 18 January 2016
Available online 16 February 2016

Keywords:
Maximum entropy
Minimum discrimination information
Joint distribution
KL distance
Demography
Household profile

ABSTRACT

Obtaining a full joint distribution from individual marginal distributions with incomplete information is a non-trivial task that continues to challenge researchers from various domains including economics, demography, and statistics. In this work, we develop a new methodology referred to as "Generalized Cross Entropy Method" (GCEM) that is aimed at addressing the issue. The objective function is proposed to be a weighted sum of divergences between joint distributions and various references. We show that the solution of the GCEM is unique and global optimal. Furthermore, we illustrate the applicability and validity of the method by utilizing it to recover the joint distribution of a household profile of a given administrative region. In particular, we estimate the joint distribution of the household size, household dwelling type, and household home ownership in Singapore. Results show a high-accuracy estimation of the full joint distribution of the household profile under study. Finally, the impact of constraints and weight on the estimation of joint distribution is explored.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Understanding household composition and profile of a geographic area is a burgeoning topic of research in the field of demography and crucial in many aspects of urban planning [1–3]. Knowing, for example, the historical spatiotemporal

* Corresponding author.
E-mail addresses: xuh@ihpc.a-star.edu.sg (H.-Y. Xu), monterolac@ihpc.a-star.edu.sg (C.P. Monterola).

http://dx.doi.org/10.1016/j.physa.2016.02.023
0378-4371/© 2016 Elsevier B.V. All rights reserved.

trends in the demand for housing may allow city planners to gauge and project future demands, which not only affect the required number of housing units but also the demands for other amenities such as schools, parks, roads, and public transport [4].

The availability of census data has played a huge role in providing government agencies and other stakeholders with rich information about the socioeconomic reports of individuals in a population. This allows them to project populations and map out other provisions within an administrative region. On the other hand, these data are usually provided in aggregates and under distinct and separate variables due to confidentiality reasons, and/or the large quantities of data involved [5]. In order to know the household profile and composition, which is a multi-dimensional matter involving an amalgam of various measures, such as household size, dwelling type, and ownership, one needs to recover the joint distribution of the quantities under consideration.

The recovery of a joint distribution from incomplete information is a non-trivial task, and a key problem confronting many statistical researchers [6–12]. The main difficulty lies in how to incorporate all relevant information and guarantee the accuracy of the estimation. Methods like Bayesian [7,13], least square [6], and generalized method of moment [8], have been proposed for estimating a full joint distribution with known low-order joint probability, and other related information. However, Bayesian method is limited to estimate joint distributions with relatively few free cells [8], while the least squares method may lead to negative probability estimates [9]. On the other hand, the maximum entropy method [9,10,14–16] demonstrates good accuracy and appropriately describes the dynamics of various statistical and physical systems [17–21].

In this work, we develop a Generalized Cross Entropy Method (GCEM) to estimate a full joint distribution of a household profile by formulating the household size (HS), household dwelling type (HD), and home ownership (HO) measures as constraints, with the objective of minimizing a weighted sum of divergences between the estimate and a set of references. The term GCEM stems from the objective function expressed as a weighted sum of cross entropy (CE) [22,23] and entropy. The CE for discrete systems are described in Ref. [22], while for continuous systems it is reported in Ref. [23]. We show that our proposed procedure is more general and flexible as it is built to incorporate multiple references. That is, existing maximum entropy procedures [9,10,14–16] can be formulated as special cases of our framework. For purpose of illustration, we use data from the Singapore Department of Statistics (SDOS) (http://www.singstat.gov.sg) and show that GCEM yields high-accuracy estimation of the full joint distribution of the household profile.

The article proceeds as follows. In Section 2, we present the theoretical framework of GCEM. In Section 3, we illustrate the efficacy and accuracy of GCEM in dealing with Singapore household data. This is then followed by a discussion on the selection of constraints and weights in Section 4. Finally, summary and conclusions are provided in Section 5.

2. Generalized cross entropy method

2.1. Theoretical framework

Existing maximum-entropy methods are typically used for the recovery of joint distributions aimed at solving a well-defined problem described by a known reference distribution. In the absence of *a priori* estimation of the joint distribution, the reference is normally a uniform distribution (i.e., pure maximum entropy) [10]. The reference can be the product of the marginal distributions [14], if the marginal is available, or the prior estimator of a joint distribution [15,16]. However, the reference may not be the sole factor that needs to be considered as there may exist more than one prior estimate, which can be in conflict with each other. To address this limitation, we propose here a Generalized Cross Entropy Method (GCEM) where the objective function to be minimized is a weighted sum of divergences between the joint distribution and the references. That is,

$$\begin{aligned} \text{Min } E(\mathbf{p}) \quad &= \sum_i^I \omega_i E^{MDI}(\mathbf{p}, \mathbf{q}_i) + \left(1 - \sum_i^I \omega_i\right) E^{ME}(\mathbf{p}) \\ \text{s.t.} \quad & A\mathbf{p} = \mathbf{a}; \\ & B\mathbf{p} - \mathbf{b} \le 0; \\ & -p_j \le 0; \quad i = 1, \ldots, I, j = 1, \ldots, J \end{aligned} \tag{1}$$

where $\mathbf{p} = [p_1, \ldots, p_J]^\top$ is the joint distribution with J representing the dimension; $\mathbf{q}_i = [q_{i1}, \ldots, q_{ij}]^\top$ is an initial estimate of $\mathbf{p}$; I is the number of references; $\omega_i \ge 0$ ($\sum_i \omega_i \le 1$) is the weight on $\mathbf{q_i}$, or the researcher's belief on $\mathbf{q_i}$; A and B are matrices, $\mathbf{a}$ and $\mathbf{b}$ are vectors, where $A\mathbf{p} = \mathbf{a}$ and $B\mathbf{p} - \mathbf{b} \le 0$ represent the constraint conditions of the model.

The first term in Eq. (1) is a weighted sum of the Kullback–Leibler (KL) distance [24] between the joint distribution and its prior estimates. Minimizing this term follows the minimum discrimination information (MDI) principle. This part is thus denoted by $E^{\mathrm{MDI}}(\mathbf{p}, \mathbf{q}_i)$. The second term is a proportion of a negative entropy, i.e., the distance between the joint distribution and a uniform distribution. To minimize this part, we follow the principle of maximum entropy (ME)[25] denoted by $E^{\mathrm{ME}}(\mathbf{p})$.

Download English Version:

https://daneshyari.com/en/article/973619

Download Persian Version:

https://daneshyari.com/article/973619

Daneshyari.com