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.



## 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/

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 $E^{\mathrm{ME}}(\mathbf{p})$.

**Remark 1.** The objective function in Eq. (1) simplifies to

$$\begin{aligned} E(\mathbf{p}) &= \sum_i \sum_j \omega_i p_j \log(p_j/q_{ij}) + \left(1 - \sum_i \omega_i\right) \sum_j p_j \log(p_j) \\ &= -\sum_i \sum_j \omega_i p_j \log(q_{ij}) + \sum_j p_j \log(p_j) \\ &= \sum_i H(\mathbf{p}, \mathbf{q}_i) - H(\mathbf{p}). \end{aligned} \tag{2}$$

As Eq. (2) is actually a weighted sum of cross entropy [22] $H(\mathbf{p}, \mathbf{q}_i)$ and entropy $H(\mathbf{p})$, we refer this method as GCEM.

**Remark 2.** The existing maximum entropy objective functions are found to be special cases of our objective functions. If $\sum_i \omega_i = 0$, the objective function in Eq. (1) is a pure maximum entropy (see Ref. [10]). If $\sum_i \omega_i = 1$, the objective function becomes the minimum discrimination information. This objective function is identical to what has been described in the maximum entropy weights imputation procedure [9,15,16]. If the reference distribution is a product of marginal distributions, it becomes [14]. This illustrates that the proposed GCEM is more general and flexible than existing maximum entropy methods.

**Remark 3.** The objective function in the proposed GCEM can also be extended to continuous system with

$$E^{MDI}(P, Q) = \int_{-\infty}^{+\infty} (P(x) \log(P(x)/Q(x)) + (1 - Q(x)) \log((1 - Q(x))/(1 - Q(x)))\mathrm{d}x$$

and

$$E^{ME}(P) = \int_{-\infty}^{+\infty} (P(x) \log P(x) + (1 - P(x)) \log(1 - P(x)))\mathrm{d}x$$

where $P$ is the objective joint cumulative distribution function and $Q$ is the reference distribution (see Ref. [23]).

### 2.2. Model solution

The Lagrangian of Eq. (1) is given by:

$$L(\mathbf{p}; \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = E(\mathbf{p}) + \boldsymbol{\lambda}^\top (A\mathbf{p} - \mathbf{a}) + \boldsymbol{\beta}^\top (B\mathbf{p} - \mathbf{b}) - \boldsymbol{\gamma}^\top \cdot \mathbf{p} \tag{3}$$

with $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ being the Lagrange multipliers [26–28].

The Hessian matrix $H$ of the objective function in Eq. (1) is:

$$H = \frac{\partial^2 E(\mathbf{p})}{\partial \mathbf{p}^2} = \begin{bmatrix} 1/p_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/p_2 & 0 & \cdots & 0 \\ & \vdots\vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/p_J \end{bmatrix}, \tag{4}$$

where $E(\mathbf{p})$ is expressed as Eq. (2). Since $p_j \geq 0$ for all $j = 1, \ldots, J$, it is easy to see that $H$ is a definite positive matrix. As the objective function and the feasible domain of Eq. (1) are convex when $A$ and $B$ are known, hence the problem is a convex optimization. Therefore, we have the following conclusion:

**Theorem 1.** *If the optimization problem defined by Eq. (1) with known matrices A and B has a solution, that solution is unique. The unique solution of Eq. (1) can then be obtained by solving the Karush–Kuhn–Tucker (KKT) conditions [29–31], which are outlined below:*

$$\begin{cases} \dfrac{\partial E(\mathbf{p})}{\partial \mathbf{p}} + A^\top \boldsymbol{\lambda} + B^\top \boldsymbol{\beta} - \boldsymbol{\gamma} = 0; \\ A\mathbf{p} - \mathbf{a} = 0; \\ B\mathbf{p} - \mathbf{b} \leq 0; \qquad \beta_k (B_k \mathbf{p} - b_k) = 0, \quad k = 1, \ldots, K; \\ -p_j \leq 0, \qquad \gamma_j p_j = 0, \quad j = 1, \ldots, J; \\ \boldsymbol{\beta} \geq 0; \boldsymbol{\gamma} \geq 0 \end{cases} \tag{5}$$

*where $K$ is the number of rows of matrix $B$, $\beta_k$ is the $k$ th element of $\boldsymbol{\beta}$, and $B_k$ is the $k$ th row of $B$.*

**Table 1**
Singapore Household Profile. The table lists the different parameters (Dwelling Type, Household size, and Ownership) that comprise a household profile. Each of these variables is further broken down into sub types.

| $i$ | Dwelling type | $j$ | Household size | $k$ | Ownership |
|---|---|---|---|---|---|
| 1 | HDB[a] (RM[b] 1–2) | 1 | 1 person | 1 | Owner |
| 2 | HDB (RM 3) | 2 | 2 persons | 2 | Not Owner |
| 3 | HDB (RM 4) | 3 | 3 persons | | |
| 4 | HDB (RM 5+) | 4 | 4 persons | | |
| 5 | HDB (Other) | 5 | 5 persons | | |
| 6 | Condo | 6 | 6+ persons | | |
| 7 | Land | | | | |
| 8 | Other | | | | |

[a] Public flats provided by Singapore Housing & Development Board (HDB).
[b] Number of rooms.

**Proof.** For Eq. (1), its objective function and inequality constraint functions are convex, and its equality constraint functions are affine. Eq. (1) is thus a convex optimization problem (see page 136–138 in Ref. [31]). Hence, any local optimal for this optimization problem is also globally optimal. Based on the KKT conditions for convex optimization problems [29–31], the conditions for Eq. (1) are obtained immediately as shown in Eq. (5).

**Remark 4.** According to Theorem 1, a necessary condition for the existence of a unique solution of Eq. (1) is that $A$ and $B$ satisfy: the rank of matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ is $m = L + K$, where $L$ and $K$ are the number of rows of matrix $A$ and $B$, respectively, and $L + K \leq J$.

**Remark 5.** The constraint conditions in Eq. (1), which capture the marginal distribution, low-order joint distribution, moments, and other related characteristics of the joint distribution, are not necessarily linear. As shown in the proof of Theorem 1, the inequality constraint functions can be relaxed to be convex.

## 3. Generalized cross entropy method application

### 3.1. Problem formulation

In this section, we apply the GCEM in the estimation of the distribution of Singapore household profile. More particularly, we estimate the solution using the joint probability of the household dwelling type (HD), household size (HS) and home ownership (HO) measures.[1]

Singapore is a city-state and island country in Southeast Asia with 5.54 million residents and 719 square kilometer land in 2015. Majority of the residential housing developments in the city-state are presided over and established by the government. Around 85% of residents live in such public houses that have several types classified on the basis of the number of rooms (see Table 1). The remaining housing options are from private developers that include condominiums and landed properties.

We first define $p_{ijk}$ as the joint probability distribution of HD ($i$), HS ($j$) and HO ($k$), where $i = 1, \ldots, I, j = 1, \ldots, J$ and $k = 1, \ldots, K$. Here $I$ is the number of dwelling types, $J$ is the number of household size categories, and $K$ is the number of ownership category. A more detailed definition of $i$, $j$ and $k$ is provided in Table 1 with household profile data taken from the Singapore Department of Statistics (SDOS).[2] In total, there are 96 elements in the joint distribution $p_{ijk}$ to be estimated.

Our objective is to estimate the joint household distribution, i.e., the above mentioned 96 elements, for the year $t_0$. Although the full joint distribution for the year $t_0$ is not known, we do have some aggregate and other information which may be helpful in the estimation process. For example, we may know some joint household distributions for some other years instead of $t_0$. Moreover, for the year $t_0$, it is also possible for us to observe or obtain some aggregate data. Therefore, the challenge here is estimating the joint distribution when information is incomplete.

The SDOS public website provides joint distributions of HS, HD and HO in years 2000, 2005 and 2010. In addition, some aggregated data set tables are also available from the year 2000 to 2010, which are as follows:

1. Subtotal of the number of resident HD, from which we can obtain the proportion of each resident HD, which is denoted as $p_{i\cdot\cdot}$;
2. Subtotal of the number of resident HS, from which we can obtain the proportion of each resident HS, which is denoted as $p_{\cdot j\cdot}$;
3. Average resident households size for each HD, which is denoted as $e_i$;
4. Home ownership rates among resident households for each HD, which is denoted as $p_{k|i}$.

[1] The choice of variables arose from the authors' close discussion with the Singapore Urban Redevelopment Authority (URA).
[2] SDOS website: http://www.singstat.gov.sg/.

Using the historical joint distribution as the prior information of estimation, and the aggregated data as constraints, the GCEM in Eq. (1) yields:

$$\begin{aligned}
&\text{Min} \quad E(\mathbf{p}^{(t_0)}) = \sum_t \sum_{i,j,k} \omega_t p_{ijk}^{(t_0)} \log(p_{ijk}^{(t_0)}/q_{ijk}^{(t)}) + \left(1 - \sum_t \omega_t\right) \sum_{i,j,k} p_{ijk}^{(t_0)} \log p_{ijk}^{(t_0)} \\
&\text{s.t.} \quad \begin{cases} \sum_{j,k} p_{ijk}^{(t_0)} = p_{i..}^{(t_0)}; & \sum_{i,k} p_{ijk}^{(t_0)} = p_{.j.}^{(t_0)}; \\ \sum_{j} p_{ijk}^{(t_0)} = p_{k|i}^{(t_0)} \cdot p_{i..}^{(t_0)}; & \sum_{j} j \cdot p_{j|i}^{(t_0)} \le e_i^{(t_0)}; \\ \sum_{i,j,k} p_{ijk}^{(t_0)} = 1; & -p_{ijk}^{(t_0)} \le 0 \end{cases}
\end{aligned} \tag{6}$$

where $p_{j|i}^{(t_0)} = \sum_k p_{ijk}^{(t_0)}/p_{i..}^{(t_0)}$, $q_{ijk}^{(t)}$ is the joint probability in year $t$ ($t \neq t_0$), $\omega_i \ge 0$, and $0 \le \sum \omega_i \le 1$. $t$ can be 2000, 2005, or 2010, or any combination of them. If we let $x_{size}$ denote household size, then the average household size of dwelling type $i$ satisfies $e_i^{(t_0)} = \sum x_{size} \cdot p_{x_{size}|i}^{(t_0)}$. As $j = 6$ implies $x_{size} \ge 6$ according to the definition of $j$ (see Table 1), we have $\sum_j j \cdot p_{j|i}^{(t_0)} \le e_i^{(t_0)}$.

**Remark 6.** The constraints in Eq. (6) can be expressed in the same form as in Eq. (1), with

$$A = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & \overbrace{\cdots \quad \cdots}^{84} & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & \vdots & & & & \vdots & & & \vdots & & & \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \quad \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & \cdots \quad \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & \vdots & & & & \vdots & & & \vdots & & & \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \quad \cdots & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & \cdots \quad \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & \vdots & & & & \vdots & & & \vdots & & & \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \quad \cdots & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & \cdots \quad \cdots & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}_{21\times 96}, \tag{7}$$

$$\mathbf{a} = [p_{1..}^{(t_0)}, \ldots, p_{7..}^{(t_0)}, p_{.1.}^{(t_0)}, \ldots, p_{.5.}^{(t_0)}, p_{1.1}^{(t_0)}, \ldots, p_{8.1}^{(t_0)}, 1]^\top, \tag{8}$$

and

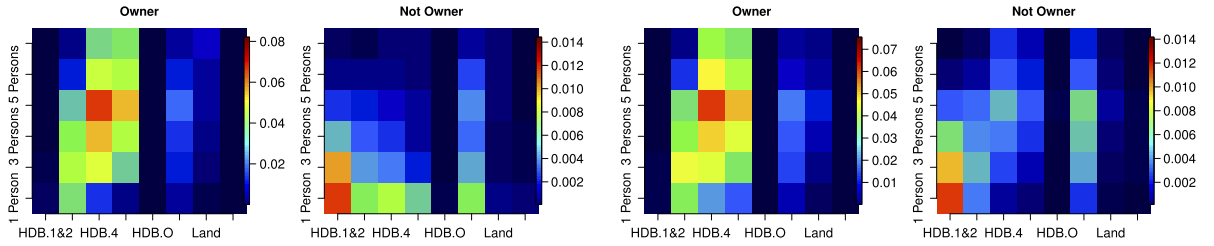
$$B = \begin{bmatrix}
1 & 1 & \cdots & 6 & 6 & \overbrace{\cdots \quad \cdots}^{84} & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & \vdots & & & \vdots & & & \vdots & & & \\
0 & 0 & 0 & 0 & 0 & \cdots \quad \cdots & 1 & 1 & \cdots & 6 & 6 &
\end{bmatrix}_{8\times 96}, \tag{9}$$

$$\mathbf{b} = [e_1^{(t_0)}, \ldots, e_8^{(t_0)}]^\top. \tag{10}$$

**Remark 7.** The design of $A$ and $B$ is to guarantee their full-rankness. The first 7 rows of matrix $A$ and the first 7 elements of vector $\mathbf{a}$ give the first 7 equations in $\sum_{j,k} p_{ijk}^{(t_0)} = p_{i..}^{(t_0)}$ ($i = 1, \ldots, 7$). When $i = 8$, $\sum_{j,k} p_{8jk}^{(t_0)} = p_{8..}^{(t_0)}$ is collinear with $\sum_{i,j,k} p_{ijk}^{(t_0)} = 1$, which is given by the last row of $A$ and last element of $\mathbf{a}$. The 8th through 12th row of $A$ represents the 5 noncollinear equations in $\sum_{i,k} p_{ijk}^{(t_0)} = p_{.j.}^{(t_0)}$ with $j = 1, \ldots, 5$. Then, the 13th through 20th row of $A$ is the 8 noncollinear equations in $\sum_j p_{ijk}^{(t_0)} = p_{k|i}^{(t_0)} \cdot p_{i..}^{(t_0)}$ with $i = 1, \ldots, 8$ and $k = 1$ as $p_{i.2} = p_{i..} - p_{i.1}$ ($i = 1, \ldots, 8$). Matrix $B$ and vector $\mathbf{b}$ reflect the constraints in $\sum_j j \cdot p_{j|i}^{(t_0)} \le e_i^{(t_0)}$, $i = 1, \ldots, 8$.

For the purpose of validation, we estimate the joint distribution for the year 2010 (i.e., $t_0 = 2010$) based on the historical joint distribution and the available aggregated data set tables.

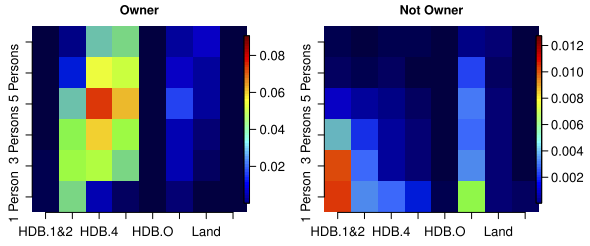
**Remark 8.** As discussed above, we do not know the full joint distribution for the year 2010. However, it is possible for us to know some aggregate information. For example, we do know certain marginal information such as $p_{i..}^{(2010)}$, $p_{.j.}^{(2010)}$, $e_i^{(2010)}$ $p_{k|i}^{(2010)}$ in Eq. (6).



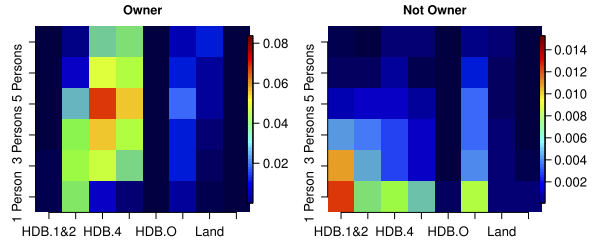

(a) True joint distribution in 2010.




(b) EM (Case 1).




(c) MDI without constraints (Case 2).




(d) MDI with constraints (Case 3).

**Fig. 1.** Accuracy Heat Map. Here we compare how accurate the 2010 estimations are for the three cases considered (b–d) based on the true 2010 joint distribution (a). Each grid represents $p_{ijk}$ values, where $i$ represents the Household Dwelling (HD) type, $j$ represents Household Size (HS), and $k$ for Home Ownership (HO).

### 3.2. Results comparison and validation

In this section, we compare three different cases of Eq. (6):

*Case* 1. Pure entropy $E^{ME}(\mathbf{p})$ with constraints shown in Eq. (6).
*Case* 2. Minimum discrimination information (MDI) $E^{MDI}(\mathbf{p}, \mathbf{q}_i)$ without constraints.
*Case* 3. MDI $E^{MDI}(\mathbf{p}, \mathbf{q}_i)$ with constraints shown in Eq. (6).

For *Case* 1, we do not have any prior information on the joint distribution. We may only maximize the entropy of $\mathbf{p}^{(\mathbf{2010})}$. Then from Eq. (6), the objective function reduces to:

$$E(\mathbf{p}^{(\mathbf{2010})}) = \sum_{i,j,k} p_{ijk}^{(2010)} \log p_{ijk}^{(2010)}. \tag{11}$$

For *Case* 2, we consider that historical joint distribution for the year 2005 is known. We can then utilize this as *a priori* estimator for $\mathbf{p}^{(\mathbf{2010})}$. If we only consider the MDI between $\mathbf{q}^{(\mathbf{2005})}$ and $\mathbf{p}^{(\mathbf{2010})}$, the objective function in Eq. (6) then becomes:

$$E(\mathbf{p}^{\mathbf{2010}}) = \sum_{i,j,k} p_{ijk}^{(2010)} \log(p_{ijk}^{(2010)}/q_{ijk}^{(2005)}). \tag{12}$$

It is easy to see that in order to minimize the above problem without any constraints, we simply set $\mathbf{q}^{(\mathbf{2005})}$ as an estimator for $\mathbf{p}^{(\mathbf{2010})}$.

For *Case* 3, the optimization problem is identical to Eq. (6) with $t = 2005$, and $\omega = 1$. The joint distribution can then be obtained through model optimization.

We then use three different measure methods to compare the performances of the above three estimates of the joint distribution. The three measures are heat map visualization, Kullback–Leibler (KL) distance and Linfoot's measures. In all these statistical measures, the procedure in *Case* 3 yields the best estimates.

Fig. 1 compares the accuracy of the three cases using heat maps. Each grid represents a $p_{ijk}$ with $i = 1, \ldots, 8, j = 1, \ldots, 6$ and $k = 1, 2$. We can visually observe that the estimate based on MDI with constraints (*Case* 3) gives the best results.

We now utilize Kullback–Leibler (KL) distance [24] to compare the estimated joint distribution and the observed one. The smaller the KL distance between any two distributions is, the closer are their profiles. The KL distance between the discrete probability distributions $P$ and $Q$, with $P$ being the real distribution, is defined as:

$$D_{KL}(P \parallel Q) = \sum_i \log\left(\frac{P(i)}{Q(i)}\right) P(i). \tag{13}$$

In order to further evaluate the estimation performance, we utilized 5000 randomly sampled joint distributions constrained by the conditions in Eq. (6). Fig. 2 shows the probability distribution of the KL distance between the real 2010 joint distribution and the 5000 randomly generated joint distributions. We find that all three estimation methods (*Case* 1–3)

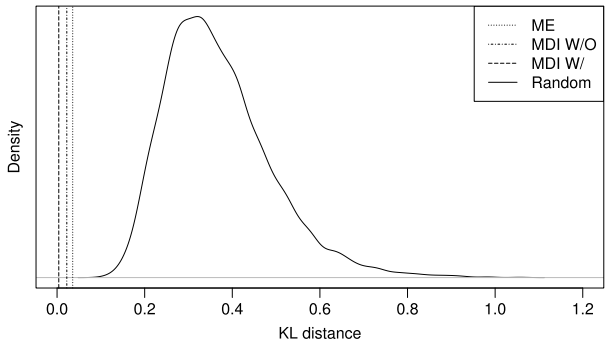



**Fig. 2.** Kullback–Leibler (KL) distance. The figure shows KL distance between the true joint distribution in 2010 and its estimation based on the cases: ME (*Case* 1, dotted), MDI without constraints (*Case* 2, dotdash), and MDI with constraints (*Case* 3, longdash). The curve (solid line) shows the probability density of KL distance between the true joint distribution in 2010 and 5000 randomly sampled joint distributions. Lower K–L distances essentially lead to more accurate estimates.

**Table 2**
KL Distance and Linfoot's Measures. The table summarizes the accuracy measures of three estimates (ME, MDI with constraints, and MDI without constraints) of the 2010 joint distribution.

| Measures | | Estimates | | | |
|---|---|---|---|---|---|
| | | ME (Case 1) | MDI W/O (Case 2) | MDI W/ (Case 3) | 5000 samples (Mean ± SD) |
| KL distance | | 0.0358 | 0.0223 | 0.0039 | 0.3688 ± 0.1215 |
| Linfoot's measure | $C$ | 0.940* | 1.144** | 1.016 | 1.4164 ± 0.1696** |
| | $Q$ | 0.987 | 0.988 | 0.999 | 0.6263 ± 0.1578** |
| | $F$ | 0.962 | 1.066* | 1.007 | 1.0214 ± 0.0230 |

* $0.05 \leq |x-1| < 0.1$.
** $0.1 \leq |x-1|$.

perform much better than a randomly selected joint distribution (about 1–2 orders of magnitude improvement) while the estimate based on MDI with constraints performs the best. KL distance between the three estimates to the observed probability is listed in Table 2.

Finally, we use Linfoot's measures defined as:

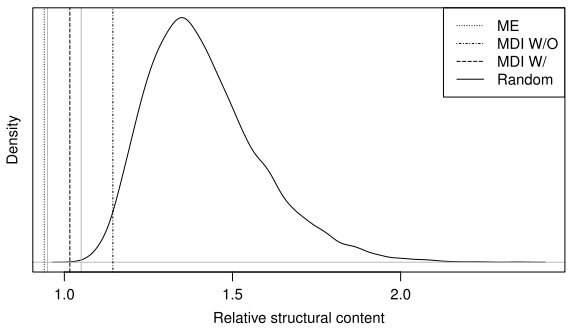
$$\begin{cases} C = \dfrac{\sum_i Q^2(i)}{\sum_i P^2(i)} \\ F = 1 - \dfrac{\sum_i (P(i) - Q(i))^2}{\sum_i P^2(i)} \\ Q = \dfrac{\sum_i P(i) * Q(i)}{\sum_i P^2(i)}, \end{cases} \tag{14}$$

where $C$ measures the relative structural content, $F$ looks at the fidelity or peak alignment, and $Q$ reflects the correlation quality. Linfoot's criterion has been previously used to compare a wide range of spatiotemporal signals from brain waves to human dynamics [12,32,33]. If $C = F = Q = 1$, $P$ is identical to $Q$. If all the values are within $(0.95, 1.05)$, $P$ has a strong statistical match with $Q$. Linfoot's measures between estimated joint distribution and observed distribution are shown in Table 2. The distributions of the Linfoot's measures between the 5000 randomly sampled joint distributions and the observed one (2010) are shown in Fig. 3.

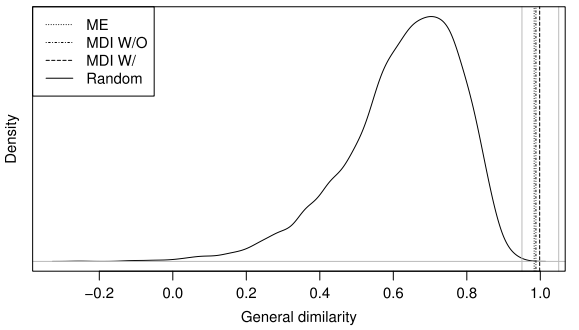
Consistent with previous results, the three estimated joint distributions perform much better than a randomly sampled joint distribution. The performance of the estimates based on pure ME (*Case* 1) and the estimates directly obtained from MDI without constraints (*Case* 2) are comparable to each other, while the estimation based on MDI with constraints (*Case* 3) is the most accurate.

## 4. Effect of constraints and weight in the accuracy of the estimate

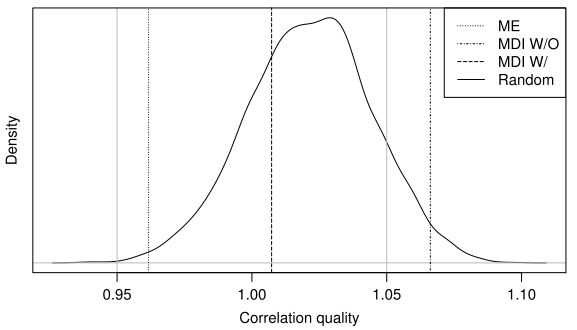
We now explore how constraints and weight affect the estimation of joint distribution. The analysis is useful to guide the selection of the two parameters in Eq. (1).




(a) Relative structural content.




(b) General similarity.




(c) Correlation quality.

**Fig. 3.** Linfoot's Criteria. The figure shows Linfoot's measures (a) relative structural content, (b) general similarity, and (c) correlation quality, between the true joint distribution in 2010 and its estimation based on the cases: ME (*Case* 1, dotted), MDI without constraints (*Case* 2, dotdash), and MDI with constraints (*Case* 3, longdash). The curves (solid line) show the probability densities of the Linfoot's criteria for 5000 randomly sampled joint distributions. For measures closer to 1.0, they are considered more accurate.

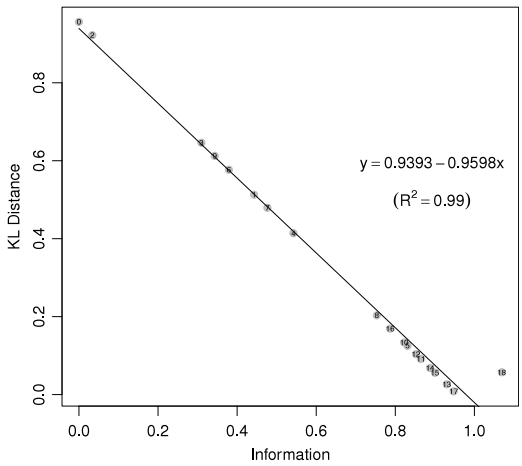
### 4.1. Effect of constraints

We use *Case* 1 and *Case* 3 described in the previous section, and play with various constraints. Table 3 lists all the possible constraints, consisting of all low-order joint distributions and marginal distributions, of the Singapore household profile data. As in previous representation, $i$, $j$ and $k$ denote HD, HS and HO, respectively. The corresponding joint distributions of each of the two variables are denoted as $p_{ij.}$, $p_{i.k}$ and $p_{.jk}$. The corresponding marginal distributions are $p_{i..}$, $p_{.j.}$ and $p_{..k}$. Since the information contained in the joint distribution of two variables also includes the information contained in their marginal distribution, there are 18 groups of constraints in total, denoted as 0–17 as shown in Table 3. To calculate the information contained by each constraint group, we first estimate the joint distribution determined purely by the constraint, i.e., pure maximum entropy method described in *Case 1*. Then, the information is measured by the KL distance between the estimated joint distribution and the uniform distribution. The idea behind this measurement is that the bigger the distance from a specified distribution is to a uniform distribution, the more information it contains. For constraint group 0, no information is provided. On the other hand, constraint group 18 is the constraint group described in Eq. (6).

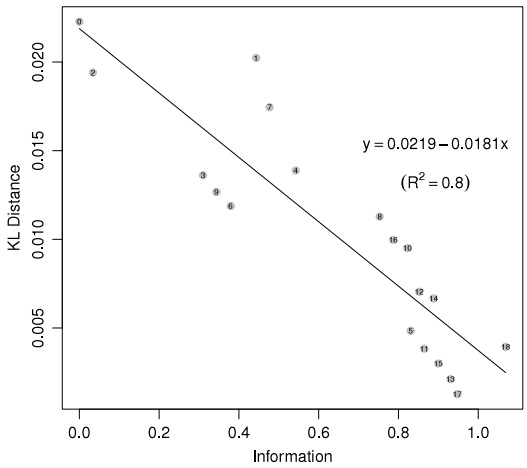
**Remark 9.** In a constraint group, if all the constraints are distributions and there are no interaction among the constraints, the information described above is the sum of the KL distance between all of the constraints and their corresponding uniform distributions. For instance, in constraint group 7 (Table 3), the constraints are $p_{i..}(i = 1, \ldots, 8)$ and $p_{.j.}$ $(j = 1, \ldots, 6)$. The information is calculated as

$$\sum_i p_{i..} \log(p_{i..}/(1/8)) + \sum_j p_{.j.} \log(p_{.j.}/(1/6)) = 0.4769.$$

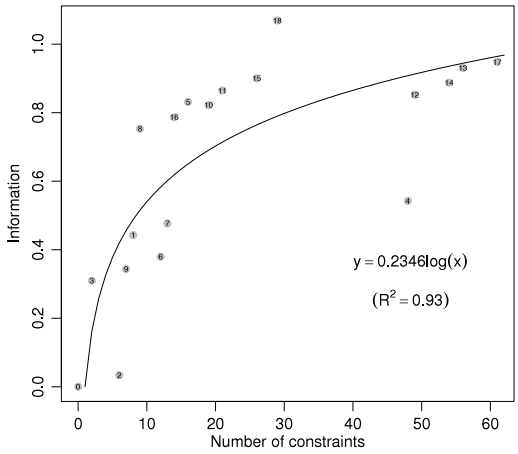
We again use KL distance between the estimated joint distribution and the real distribution to measure the accuracy of estimates (see Fig. 4(a) and (b)). It is observed that the accuracy of estimation is highly determined by the magnitude of information contained in the constraints. In *Case* 1, $R^2$ of regressing estimates' accuracy on constraints' information is 0.99, while in *Case* 3, $R^2 = 0.8$. In Fig. 4(c), the relationship between the information contained in each constraint group and the number of constraints is observed to be described by a logarithm function with $R^2 = 0.93$. The selection of a proper constraint group can be conducted according to Fig. 4(c). The points that locate above the fitted curve have a higher capability




(a) *Case* 1 (ME).




(b) *Case* 3 (MDI).



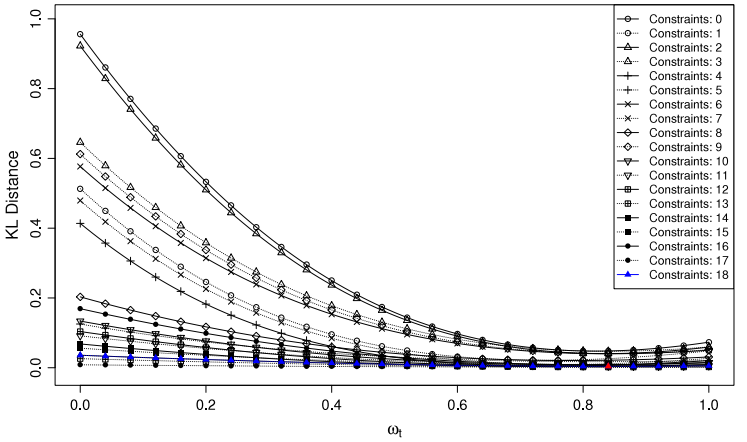

(c) Information vs. Number of constraints.

**Fig. 4.** Estimation Error, Information and Number of Constraints. Figure (a) and (b) show the plots of estimation error (represented by the KL distance between the estimated joint distribution and the observed joint distribution) vs. the information of constraints (represented by the KL distance between the constraint and its corresponding uniform distributions); Figure (c) shows the plots of information of each constraint group (represented by the KL distance between the estimated joint distribution and the uniform distribution) as a function of the number of constraints. The constraint group indices shown in Table 3 are labeled in the plots here accordingly.

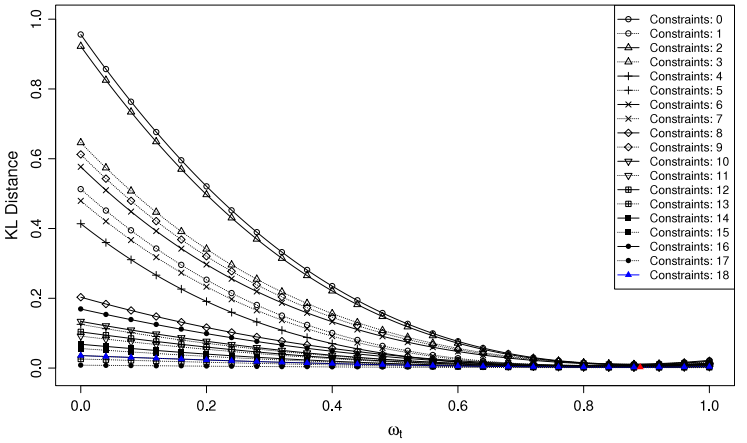
**Table 3**
List of constraints.

| Index | Constraints | Information | #(Constraints) | Index | Constraints | Information | #(Constraints) |
|---|---|---|---|---|---|---|---|
| 0 | – | 0 | 0 | 10 | $p_{i..}, p_{.jk}$ | 0.8224 | 19 |
| 1 | $p_{i..}$ | 0.4431 | 8 | 11 | $p_{.j.}, p_{i.k}$ | 0.8646 | 21 |
| 2 | $p_{.j.}$ | 0.0338 | 6 | 12 | $p_{..k}, p_{ij.}$ | 0.8524 | 49 |
| 3 | $p_{..k}$ | 0.3100 | 2 | 13 | $p_{ij.}, p_{i.k}$ | 0.9302 | 56 |
| 4 | $p_{ij.}$ | 0.5424 | 48 | 14 | $p_{ij.}, p_{.jk}$ | 0.8879 | 54 |
| 5 | $p_{i.k}$ | 0.8308 | 16 | 15 | $p_{i.k}, p_{.jk}$ | 0.9002 | 26 |
| 6 | $p_{.jk}$ | 0.3793 | 12 | 16 | $p_{i..}, p_{.j.}, p_{..k}$ | 0.7868 | 14 |
| 7 | $p_{i..}, p_{.j.}$ | 0.4769 | 13 | 17 | $p_{ij.}, p_{i.k}, p_{.jk}$ | 0.9478 | 61 |
| 8 | $p_{i..}, p_{..k}$ | 0.7530 | 9 | 18 | $p_{.j.}, p_{i.k}, e_i$ (Eq. (6)) | 1.0689 | 29 |
| 9 | $p_{.j.}, p_{..k}$ | 0.34387 | 7 | | | | |

of contained information than the points below the curve. Hence, constraint group numbers 3, 5, 8, 10, 11, 15, 16 and 18 are more efficient than other groups. Meanwhile, constraint group 18 recommended by SDOS represents a good compromise between the number of constraints required and the information contained.




(a) $t = 2000$.




(b) $t = 2005$.

**Fig. 5.** Estimation Error vs. Weight. The figure shows the plots of estimation error (represented by the KL distance between the estimated joint distribution and the observed joint distribution) as a function of the weight considered in the joint distribution estimation for both the joint distributions of (a) 2000 and (b) 2005 as references for various constraint groups (see Table 3).

### 4.2. Effect of weight

To see the impact of weight, Eq. (6) is re-written as:

$$E(\mathbf{p}^{(t_0)}) = \sum_{i,j,k} p_{ijk}^{(t_0)} (\log p_{ijk}^{(t_0)} - \omega_t \log q_{ijk}^{(t)}) \tag{15}$$

where $t_0 = 2010$, and $t = 2005$. When $\omega_t = 0$, it is the same as Eq. (11); while when $\omega_t = 1$, it is the same as Eq. (12). The plot of KL distance between the estimated joint distribution under each constraint group (Table 3) and the observed distribution (2010) with respect to the weight $\omega_t$ is shown in Fig. 5(b). Whereas, the estimation based on the joint distribution at $t = 2000$ as the prior estimate is also conducted to compare the impact of weight when prior estimate has a different accuracy. The corresponding plot of KL distance with respect to the weight for the case $t = 2000$ is given in Fig. 5(a). The KL distance between $\mathbf{q}^{(2000)}$ and $\mathbf{p}^{(2010)}$ is 0.0732, while the KL distance between $\mathbf{q}^{(2005)}$ and $\mathbf{p}^{(2010)}$ is 0.0223, suggesting that $\mathbf{q}^{(2005)}$ is closer to $\mathbf{p}^{(2010)}$.

In addition, Fig. 5 indicates that KL distance is convex and non-monotonic with respect to weight. The optimal weight on $\mathbf{q}^{(2005)}$ is higher than that on $\mathbf{q}^{(2000)}$. For example, if we employ constraint group 18 (see Table 3), the optimal $\omega_{2005}$ is 0.89, while for $\omega_{2000}$ is 0.84. It suggests that higher weight should be assigned closer to the objective joint distribution.

To determine the weight, it would be helpful to incorporate recommendations from experts in the specific area. Nevertheless, it is observed from Fig. 5 that the impact of weight becomes weaker and weaker as the information provided

by the constraint group increases. As a case in point, when the information contained in the constraint group is more than 0.8,[3] the curve of estimation accuracy as a function of weight is flat, which exhibits a weak impact of the weights.

## 5. Conclusions

We have proposed a Generalized Maximum Entropy Method (GCEM) to estimate full joint distributions from incomplete information. The objective function was formulated using a weighted sum of the Kullback–Leibler (KL) distance between the joint distribution and the references. The constraints were formed from marginal distributions, low-order joint distributions, moments and other related characteristics of the joint distribution. The accuracy of the method was illustrated by estimating a joint distribution of Singapore household with respect to household dwelling type (HD), household size (HS), and household home ownership (HO). Our method yielded a KL distance that was about an order of magnitude better than the standard maximum entropy method. We also described a method to measure the information contained in each constraint group that was shown to improve the estimation accuracy. Finally, we outlined how our developed procedure is applicable to many other areas of research where the recovery of joint distributions from incomplete information is critical.

## Acknowledgments

This research was supported by Singapore A*STAR SERC Complex Systems Programme research grant (1224504056 for SK, EL, CM) and Integrated City Planning research grant (1325000001 for HX, SK, GL, DZ).

## References

[1] G. Bramley, H. Pawson, M. White, D. Watkins, N. Pleace, Estimating Housing Need. Department for Communities and Local Government, Eland House, Bressenden Place, London, 2010, www.communities.gov.uk.
[2] J.J. Salvo, W.A. Brown, Population Estimates and the Needs of Local Governments Paper Presented at US. Census Bureau Conference on Population Estimates: Meeting User Needs, Alexandria VA, 2006.
[3] D.A. Swanson, G.C. Hough Jr., An evaluation of persons per household (PPH) estimates generated by the American community survey: A demographic perspective, Population 31 (2012) 235–266.
[4] S.K. Smith, J. Nogle, S. Cody, A regression approach to estimating the average number of persons per household, Demography 39 (2002) 697–712.
[5] J.-J. Salazar-Gonzalez, Statistical confidentiality: Optimization techniques to protect tables, Comput. Oper. Res. 35 (2008) 1638–1651.
[6] I. Csiszar, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, Ann. Statist. 19 (1991) 2032–2066.
[7] D.S. Putler, K. Kalyanam, J.S. Hodges, A Bayesian approach for estimating target market potential with limited geodemographic information, J. Mar. Res. 33 (1996) 134–149.
[8] C.J. Romeo, Estimating discrete joint probability distributions for demographic characteristics at the store level given store level marginal distributions and a city-wide joint distribution, QME-Quant. Mark. Econ. 3 (2005) 71–93.
[9] M. Ruther, G. Maclaurin, S. Leyk, B. Buttenfield, N. Nagle, Validation of spatially allocated small area estimates for 1880 Census demography, Demogr. Res. 29 (2013) 579–616. http://dx.doi.org/10.4054/DemRes.2013.29.22.
[10] A.E. Abbas, Entropy methods for joint distributions in decision analysis, IEEE Trans. Eng. Manag. 53 (2006) 146–159.
[11] G. Li, D. Zhao, Y. Xu, S.-H. Kuo, H.-Y. Xu, N. Hu, G. Zhao, C. Monterola, Entropy based modelling for estimating demographic trends, PLoS One 10 (9) (2015) e0137324.
[12] C. Monterola, M. Lim, J. Garcia, C. Saloma, Accurate forecasting of undecided population in a public opinion poll, J. Forecast. 21 (2002) 435–449.
[13] Y.P. Chaubey, F. Nebebe, K. Dzicielowski, D. Sen, Estimation of joint distribution from marginal distributions, in: Proceedings of the American Statistical Association, Section on Survey Research Methods, 2003, pp. 883–888.
[14] D.J. Miller, W. Liu, On the recovery of joint distributions from limited information, J. Econometrics 107 (2002) 259–274.
[15] S. Leyk, B.P. Buttenfield, N.N. Nagle, Modeling ambiguity in census microdata allocations to improve demographic small area estimates, Trans. GIS 7 (2013) 406–425.
[16] S. Leyk, N.N. Nagle, B.P. Buttenfield, Maximum entropy dasymetric modeling for demographic small area estimation, Geogr. Anal. 45 (2013) 285–306.
[17] J.E. Contreras-Reyes, R.B. Arellano-Valle, Kullback–Leibler divergence measure for Multivariate Skew- Normal distributions, Entropy 14 (2012) 1606–1626.
[18] D.F.B. Haeufle, M. Gnther, G. Wunner, S. Schmitt, Quantifying control effort of biological and technical movements: An information-entropy-based approach, Phys. Rev. E 89 (2014) 012716.
[19] S. Stramaglia, G.-R. Wu, M. Pellicoro, D. Marinazzo, Expanding the transfer entropy to identify information circuits in complex systems, Phys. Rev. E 86 (2012).
[20] J.E. Contreras-Reyes, Asymptotic form of the Kullback–Leibler divergence for multivariate asymmetric heavy-tailed distributions, Physica A 395 (2014) 200–208.
[21] M.J. Salois, Regional changes in the distribution of foreign aid: An entropy approach, Physica A 392 (2013) 2893–2902.
[22] D.P.-T. Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (2005) 19–67.
[23] X. Chen, S. Kar, D.A. Ralescu, Cross-entropy measure of uncertain variables, Inform. Sci. 201 (2012) 53–60.
[24] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
[25] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. Ser. II 106 (1957) 620–630.
[26] I.B. Vapnyarskii, Lagrange multipliers, in: Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, 2001.
[27] G. Li, C. Wen, G.Z. Guo Li, A. Zhang, F. Yang, K. Mao, Model-Based Online Learning With Kernels, IEEE Trans. Neural Netw. 24 (2013) 356–369.
[28] G. Li, C. Wen, Error tolerance based support vector machine for regression, Neurocomputing 74 (2011) 771–782.
[29] H.W. Kuhn, A.W. Tucker, Nonlinear programming, in: Proceedings of 2nd Berkeley Symposium, University of California Press, Berkeley, 1951, pp. 481–492.
[30] M.A. Hanson, Invexity and the Kuhn–Tucker theorem, J. Math. Anal. Appl. 236 (1999) 594–604.
[31] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, 2008.
[32] C. Monterola, M. Zapotocky, Noise-enhanced categorization in a recurrently reconnected neural network, Phys. Rev. E (3) 71 (2005) 036134.
[33] C. Monterola, R.M. Roxas, S.L. Carreon-Monterola, Characterizing the effect of seating arrangement on classroom learning using neural networks, Complexity 14 (2009) 26–33.

[3] Information contained in constraint group 5, 10, 11, 12, 13, 14, 15, 17 and 18 is more than 0.8 according to Table 3.