扩样数据的极大熵算法

2018 年 11 月 5 日

1 介绍

这篇文章主要复现文章 [3] 的结果, 利用极大熵算法求解联合概率分布，再依据边界分布再对每一个统计量分别采样，从何合成高维数据。

2 算法介绍

对于在有限集 I 上的联合概率密度函数 p，这里仅考虑具有对数分布的联合概率密度函数

$$\boldsymbol{p} = \pi \prod_{r=1}^{c} \lambda_r^{a_{ri}} \tag{1}$$

其中 λ_r 为单个统计量概率函数，且满足如下约束等式，

$$a_r \boldsymbol{p} = h_r, \qquad r = 1, 2, \cdots, c \tag{2}$$

其中由于对数分布的函数 $\boldsymbol{p}$，

$$r_i = \begin{cases} 1, & \text{for } T = x \\ 0, & \text{else} \end{cases}$$

引入熵 $I(\boldsymbol{p}||\boldsymbol{\pi}) = \boldsymbol{p} \log \frac{\boldsymbol{p}}{\boldsymbol{\pi}}$, 其中 π 是另一个概率密度函数，而熵 $I(\boldsymbol{p}||\boldsymbol{\pi})$ 则度量了两个概率密度函数的相似程度，类似于几何距离，而极大熵

$$I(\boldsymbol{q}||\boldsymbol{\pi}) = \min_{p \in \mathcal{E}} I(\boldsymbol{p}||\boldsymbol{\pi}) \tag{3}$$

称之为 $\boldsymbol{\pi}$ 在概率空间 $\mathcal{E}$ 上的 $I-$ 投影，且 $\boldsymbol{p} = \boldsymbol{q}^*$, 其具体的几何性质在文章 [1] 中有具体的解释。主要利用极大熵 $I(\boldsymbol{q}||\boldsymbol{\pi})$ 的正交性，可以依据定理，构造迭代算法 [2] 如下，

Theorem 1. 考虑如下概率分布序列 $\{\boldsymbol{p}^{(n)}; \quad n = 0, 1, 2, \dots\}$，定义如下

$$\boldsymbol{p}^{(0)} = \boldsymbol{\pi}$$
$$\boldsymbol{p}^{(n+1)} = \boldsymbol{p}^{(n)} \prod_{r=1}^{c} \left(\frac{h_r}{h_r^{(n)}} \right)^{a_{ri}}, \qquad n = 0, 1, 2, \cdots$$

其中 $h_r^{(n)} = a_{ri}\boldsymbol{p}^{(n)}$ 为当在函数 $\boldsymbol{p}$ 下的约束等式的值。可以证明序列 $\boldsymbol{p}^{(n)}$ 一致收敛到 $\boldsymbol{q}$，使得 $I(\boldsymbol{q}||\boldsymbol{\pi}) = \min_{p\in\mathcal{E}} I(\boldsymbol{p}||\boldsymbol{\pi})$。

于是对于统计量空间 $\mathcal{A} = \{(A)_1, (A)_2, \cdots, (A)_q\}$ 以及每个统计量的样本空间 $\mathcal{R}(\boldsymbol{A}_i) = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_{k_i}^{(i)}\}$ 以及全样本空间 $\boldsymbol{S} = \times_{i=1}^{q}\mathcal{R}(\boldsymbol{A}_i)$，考虑高维统计量 $T = ((A)_1, (A)_2, \cdots, (A)_q)$，并令 $X = (A_i|A_i \in C, C \subset \mathcal{A})$ 为约束变量空间，称为模式。如 $X = (A_1)$ 时，为 $\boldsymbol{p}$ 在单变量 A_1 下的变量空间；如 $X = (A_1, A_2)$ 时，为 $\boldsymbol{p}$ 在双变量 A_1, A_2 下的变量空间。而 $S_x = \times_{A_i\in C}\mathcal{R}(A_i)$ 为约束样本空间。对于给定采样目录数据，得到后验概率 $\tilde{\boldsymbol{p}}(T = x|D)$，于是给出利用极大熵拟合后验概率求解联合概率分布的算法如下。

Algorithm 1: 极大熵求解联合概率分布算法

```
Result: A joint probability distribution
Initialize p;
while p is not convergence do
    for X_i ∈ X do
        for x_{i,j} ∈ S_{X_i} and p̃(T = x_{i,j}|D) ∈ p̃ do
            compute p(T = x_{i,j});
            λ_{i,j} ← λ_{i,j} p̃(T=x_{i,j}|D) / p(T=x_{i,j});
        end
    end
end
```

利用极大熵逼近约束概率条件解空间的正交性，可以将子变量空间的分布进行分别计算，得到一个高维的约束在后续的计算中进行利用，例如可以将三个 1 维边间概率分布进行极大熵拟合得到一个 3 维的边界概率约束。因而具有分布式计算的特性。特别地，如果两组模式 X_1, X_2 为互为独立变量且 $X_1 \cup X_2 = X$，那么可以分别求得 $\boldsymbol{p}_1^*, \boldsymbol{p}_2^*$ 使得 $\boldsymbol{p}^* = \boldsymbol{p}_1^*\boldsymbol{p}_2^*$，但真实世界中人口数据统计变量很难做到完全独立，所以这一点应用的场景不多。除此之外，如果对于约束条件中较多存在 $\boldsymbol{p}(T = x) = 0$ 可以将其作为根节点，

再将其子节点的约束分别计算最终整合成一个更高维的联合概率分布。这些在原论文中都有提到。

最后对于求得的概率分布函数再依据其边际函数的分布，进行采样得到高维数据，算法如下

Algorithm 2: 合成数据采样算法

```
Initialize T = ∅;
for A_i ∈ A do
    for a_j^(i) ∈ R(A_i) do
        compute the conditional probability p*(a_j^(i)|T);
    end
    T(A_i) ← Sample(R(A_i, {p*(a_j^(i)|T)}));
end
```

3 数值算例

接下来利用人口数据去除缺省值，并分别按 5%,10% 进行抽样来算法的验证，主要验证算法 1，取 income,age,id,gender,econActivity 共 5 个类别，各个统计量的样本个数为 13,17,11,2,13。为了评价算法结果的近似程度，文章给出类两个度量，非别是

$$\mathrm{BIC}_{\mathcal{X}} = -2\log\mathcal{L}_{\mathcal{X}} + \mathcal{N}\cdot\log|D|,$$

其中

$$\log\mathcal{L}_{\mathcal{X}} = \sum_{T\in D}\log\boldsymbol{p}^*(T) = |D|\left(\sum_{X_i\in\mathcal{X}}\sum_{x_{i,j}\in S_{X_i}}\tilde{\boldsymbol{p}}(T=x_{i,j}|D)\cdot\log u_{i,j}\right),$$

$\mathcal{N}$ 为参数个数,$|D|$ 为数据的类别数,即统计量个数,以及 Kullback-Leibler(KL) divergence

$$h(\alpha,\beta) = \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}.$$

3.1 算例 1.

在算例 1. 中，对总样本的 5% 进行抽样，取各个统计量的边际密度函数利用算法 1，再分别求得联合分布的边际密度函数如下，最终得到结果如

图 1所示，并且由于对数概率函数的假设以及本身各个统计量边际分布之间并没有矛盾或者冲突，例如给定两个变量的边际分布，又给定改两个变量的二维边际分布，可能存在约束本身不符合概率乘法公式，使得考虑各个情况下的最佳逼近。因此在这个例子下算法 1仅需要 1 步就得到收敛解，

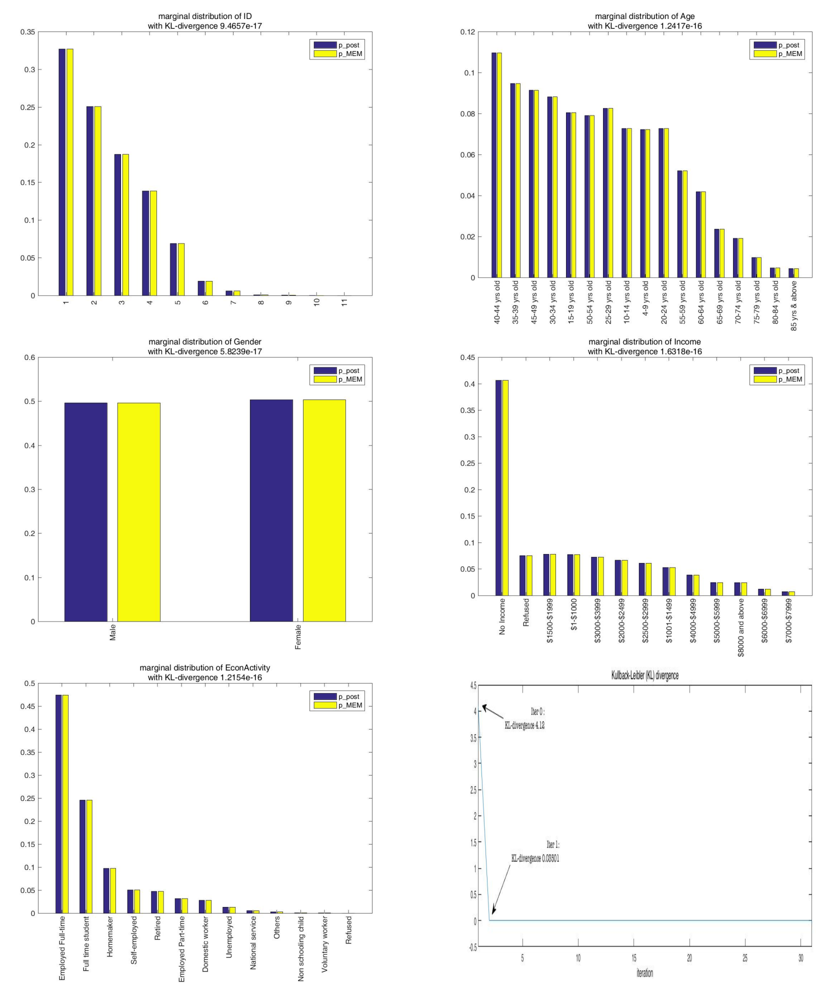

图 1: 抽样边际密度和极大熵拟合联合概率分布的边际分布对比以及 KL-divergence 迭代情况

而对比原本的全样本边际密度分布可得算法的拟合结果如下，

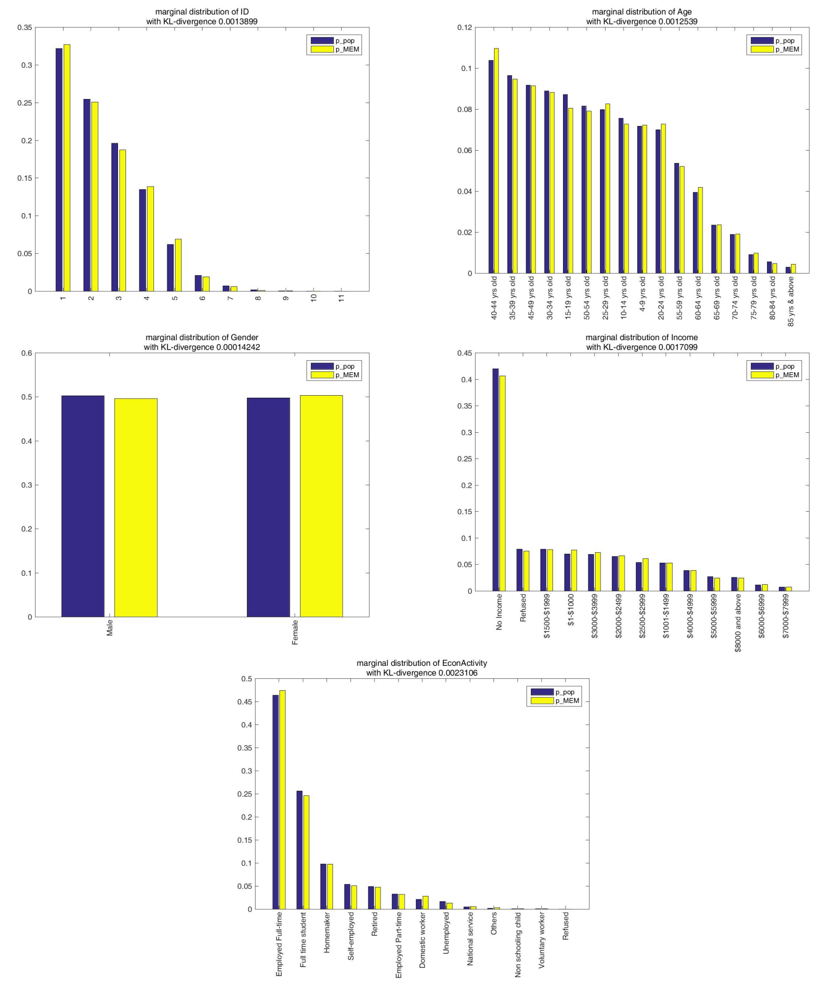

图 2: 全样本边际密度和极大熵拟合联合概率分布的边际分布对比

3.2 算例 2.

在算例 2. 中，采用对总样本 10% 进行抽样，同样可得全样本的边际分布和抽样边际分布逼近的联合概率的边际分布对比，如图 3所示。可以发现每个统计量的分布 KL-divergence 均比起 5% 的抽样结果并没有改进反而更大了。这是因为算法本身及其依赖抽样的正确性，而增大抽样概率能得到更好的对于全样本的描述，有理由相信，更大的抽样概率更能代表整体，因而也可以更好的对联合分布的拟合进行扩样，但是实际情况因人而异。

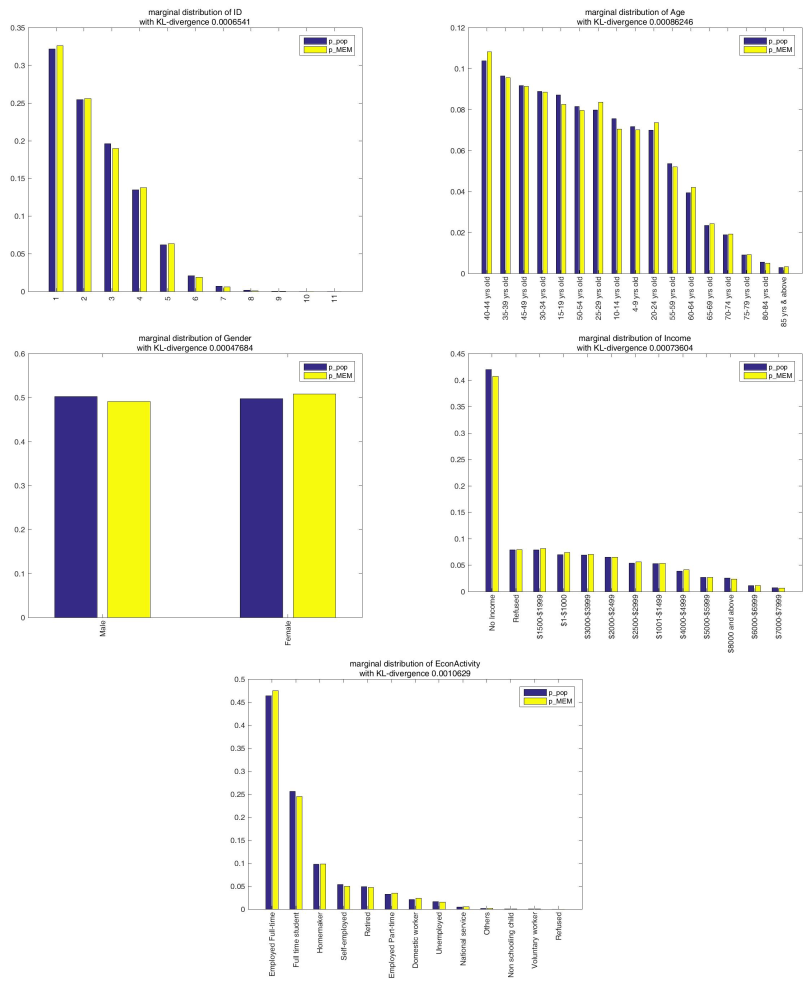

图 3: 全样本边际密度和极大熵拟合联合概率分布的边际分布对比

3.3 算例 3.

在算例 3. 中，将利用算例 1. 中得到联合概率分布中，对二维边际分布进行部分的随机采样，采取如下的统计量模式 {'ID','Age'}, {'Gender', 'Income'},{'Income','EconActivity'}，分别总共有 11*17,2*13,13*13 种组合。而部分随机采样个数采取边际分布的最小维度，和最大维度，得到均得到不错的近似结果，这里仅展示最小维度的情况如图 5。同时又将这两种方法和全部统计量的边际分布迭代的收敛情况，作对比，显然由于信息的缺失，收敛的速度会降低一些，而且由于抽样有随机性，并不能较好的覆盖尽可能多的统计变量的信息，因而部分随机抽样的迭代效果并没有随抽样个数的多少呈现较强的相关性，这些可以从图 4可以看出。并且约束个数的不同，计算量也不尽相同，算法 1所花费的时间也不同，如下表所示

表 1: 不同约束个数下的迭代时长花费

约束种类	模式个数	约束参数的个数	30 次迭代时间 (s)
全变量边际分布	5	56	0.206591
2D 最小维度抽样边际	3	26	0.112209
2D 最大维度抽样边际	3	43	0.119276

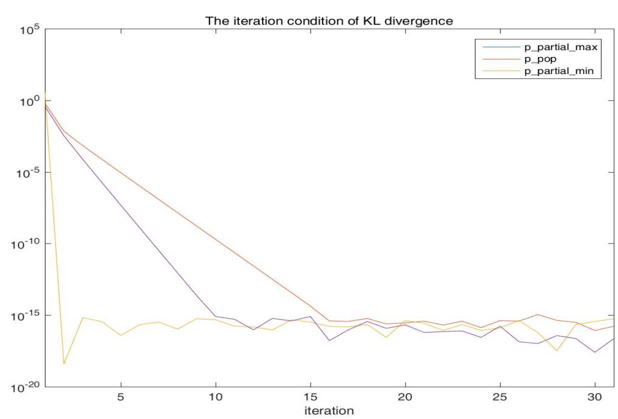

图 4: 不同约束下的 KL-divergence 迭代情况

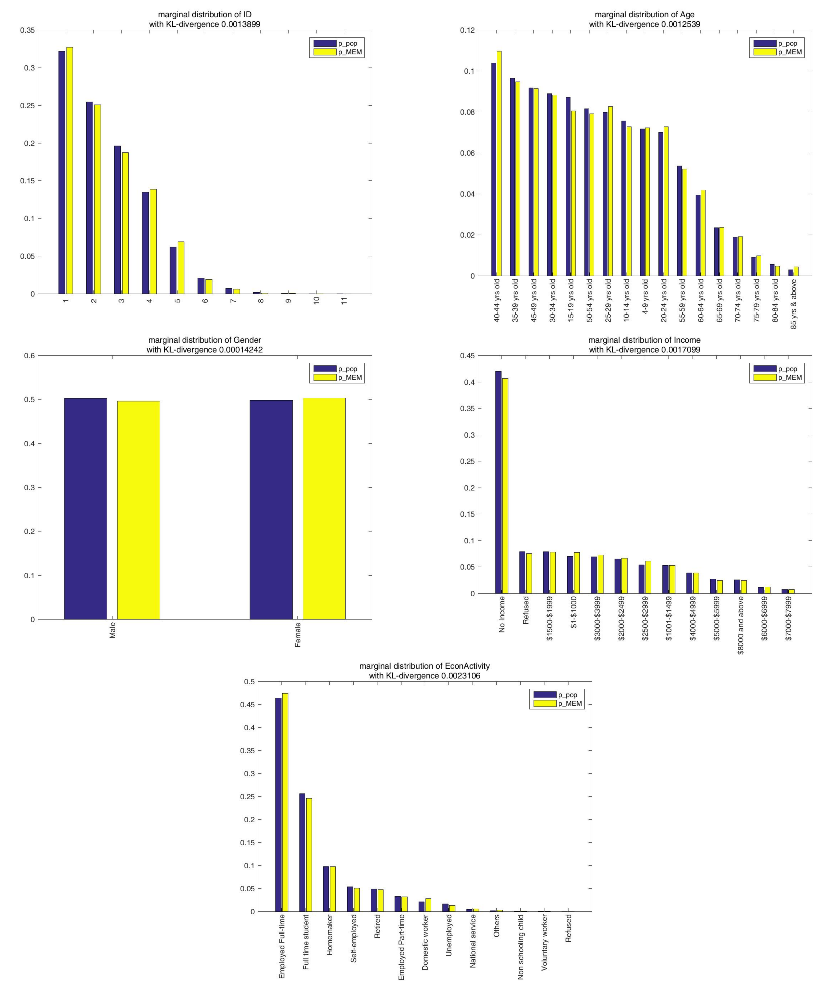

图 5: 全样本边际密度和部分抽样的极大熵拟合联合概率分布的边际分布对比

4 总结

文章中数值算例里给出的约束是二维边际分布的一部分，比如对于性别和收入两个统计量，各 2,13 个样本，一共有 26 种组合，但只给出 10 个组合的概率。像这样对于不同的统计量的组合给出不全的概率分布，又恰好可以覆盖全部变量空间（不然仅仅只是部分变量空间的概率）。并由此提出了一些基于给定约束模式的加速算法。

参考文献

[1] I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Annals of Statistics*, 11(1):141–156, 1983.

[2] J.N. Darroch Ratcliff and D. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 1(1):1–2, 1930.

[3] Hao Wu, Yue Ning, Prithwish Chakraborty, Jilles Vreeken, Nikolaj Tatti, and Naren Ramakrishnan. Generating Realistic Synthetic Population Datasets. 2016.