Keep it Simple: Addressing Rare Events in Data Synthesis Using Beta Divergence

Pavel Ilinov[*] Michel Bierlaire[†]

January 2025

1 Introduction

Generating synthetic data has gained significant importance in recent years. For instance, statistically accurate population data serves as a vital input for activity-based models, which are crucial for modeling the transportation behavior of individuals. However, due to privacy concerns, access to fully disaggregated data is often restricted. This creates a need for tools that can construct accurate population datasets using partial information from various sources.

Given the importance of this problem, it is not surprising that numerous approaches have been developed. Since synthetic population generation is essentially a statistical problem, several Machine Learning techniques have been proposed for this purpose. These techniques include Bayesian networks, variational autoencoders, and generative adversarial networks. While these methods can achieve promising results, they also exhibit significant drawbacks, such as being computationally expensive and relatively complex to implement. See Lederrey et al. (2022) for the discussion.

The classical tool for data synthesis is the so-called Iterative Proportional Fitting (IPF) algorithm. Its history dates back to at least to Deming and Stephan (1940), Ireland and Kullback (1968). Despite some limitations, the algorithm remains widely used among practitioners; see, for example, Kaddoura et al. (2024), Chang et al. (2023) for the recent application. The reason is simple: IPF is an easy-to-implement and intuitive algorithm. It can be considered a "second-best" solution that delivers sufficiently good results when simplicity and ease of implementation are prioritized.

IPF operates using two types of information: a sample dataset and aggregated data. The algorithm iteratively adjusts the joint distribution of the sample data to ensure consistency with the aggregated data. In the context of data synthesis problems, IPF is often treated as a black-box algorithm. For example, instead of modifying the algorithm itself, some researchers have proposed preprocessing the

[*] Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

[†] Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

data to mitigate its limitations Kolenikov (2014). In contrast, we take a different approach. Inspired by recent advances in the field of Machine learning, we modify the underlying optimization problem that IPF solves. Leveraging the fact that IPF is designed to solve a well-defined convex optimization problem with linear constraints, we reformulate this problem to address its drawbacks.

One of the main issues with IPF is the so-called "zero problem." Specifically, if an individual with certain characteristics is absent from the sample, they will also be absent in the synthetic data. This is clearly undesirable. Due to combinatorial effects, the zero problem is more likely to occur in small samples. A common way to address this issue is to introduce small probabilities into the empirical distribution used to generate the sample. However, this approach introduces biases in two ways. First, data manipulation can distort the properties of the empirical distribution. Second, as we demonstrate below, such manipulations bias the results obtained through IPF.

To overcome these issues, we propose a modification of the optimization problem that IPF solves. IPF traditionally seeks the distribution closest to the target distribution in terms of the Kullback-Leibler (KL) divergence. Instead, we propose using the Beta divergence, which is a one-parameter family of divergences that generalizes the KL divergence. The properties of Beta divergence have been discussed extensively in Cichocki and Amari (2010).

The use of Beta divergence is advantageous for several reasons. First, its polynomial functional form penalizes small values differently, which helps address the zero problem. Second, the KL divergence is a special case of Beta divergence when $\beta = 1$, allowing the classical solution to be retained as a specific instance. Third, Beta divergence belongs to the broader class of Bregman divergences, enabling the solution to be obtained through a simple alternating technique similar to IPF Dhillon and Tropp (2008).

Selecting the appropriate divergence measure for a statistical problem has become an increasingly active area of research, particularly in the fields of Machine Learning and Optimal Transport. For example, the classical regularized optimal transport problem can be formulated as the minimization of the KL divergence. Other divergences, such as Rényi divergence Bresch and Stein (2024) or Tsallis divergence Muzellec et al. (2017), have been applied in specific contexts for a particular application. Conceptually, our approach follows a similar rationale by exploring an alternative divergence measure to address specific shortcomings of IPF.

2 IPF

2.1 Optimization Problem

For completeness, without going into much detail, we briefly summarize the optimization problem where IPF can be applied as a numerical algorithm to find a solution.

Let there be a true distribution that we want to get an approximation of. For simplicity, our true distribution is discrete and bivariate.[1] We have two pieces of information about true distribution: the marginal distribution and a sample from it. We denote the marginal distribution as $(p_i)_{i\in 1:I}, (q_j)_{j\in 1:J}$ and sample as $(\hat{\pi}_{ij})_{i\in 1:I, j\in 1:J}$. The true distribution is $(\pi_{ij})_{i\in 1:I, j\in 1:J}$.

To approximate the true distribution, we select the closest distribution to the sample while ensuring consistency with the marginal distributions. Formally, the problem is defined as:

$$\begin{gathered}\min_{x\in\Delta^{I\times J}} D_{KL}(x;\hat{\pi})\\ \text{s.t.}\\ \sum_{j=1}^{J} x_{ij} = p_i \quad \forall i \in 1:I, \qquad (1)\\ \sum_{i=1}^{I} x_{ij} = q_j \quad \forall j \in 1:J,\end{gathered}$$

where

$$D_{KL} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}\log\frac{x_{ij}}{\hat{\pi}_{ij}}$$

is KL divergence between distributions (x_{ij}) and $(\hat{\pi}_{ij})$.

The optimality conditions can be derived from the Lagrangian formulation. Due to the infinite penalty introduced by the KL divergence for zeros, non-negativity constraints can be omitted. Taking the first-order conditions (FOC) of the Lagrangian with respect to x_{ij} and rearranging results in

$$x_{ij} = \hat{\pi}_{ij} e^{\xi_i - 1} e^{\eta_j},$$

where ξ_i and η_j are corresponding Lagrange multipliers. Denoting $u_i = e^{\xi_i - 1}$ and $v_j = e^{\eta_j}$ we get equality $x_{ij} = \hat{\pi}_{ij} u_i v_j$. Using constraints, we get the system of equations

$$p_i = u_i \sum_{j=1}^{J} \hat{\pi}_{ij} v_j, \qquad q_j = v_j \sum_{i=1}^{I} \hat{\pi}_{ij} u_i \qquad (2)$$

with respect to u_i and v_j for each $i \in 1:I$ and $j \in 1:J$. The numerical solution to the system of equations (2) is given by the IPF: the initial values of u_i and v_j are initialized and they are updated sequentially using (2). One can show that the algorithm indeed converges to the solution to the set of equations (2).

2.2 Discussion

We focus on the limitations of IPF, particularly issues related to rare events and small probability values. One such issue is the "zero problem." If an individual

[1] Below, we briefly discuss how the problem and the solution algorithms can be easily generalized for more dimensions.

with certain characteristics is absent from the sample, they will also be absent in the synthetic data. This arises due to the log penalty. Specifically, the fraction log penalty. The fraction $\log \frac{x_{ij}}{0}$ is unbounded below if $x_{ij} \neq 0$.

There is an additional implicit feature of the problem (1). Consider the situation in which $\hat{\pi}_{ij}$ is very low. In this case, value of $\log \frac{x_{ij}}{\hat{\pi}_{ij}}$ will be relatively low only if x_{ij} is the same order of $\hat{\pi}_{ij}$. If $x_{ij} >> \hat{\pi}_{ij}$ then the fraction $\frac{x_{ij}}{\hat{\pi}_{ij}}$ takes the high value. Therefore, intuitively, if $\hat{\pi}_{ij}$ is very low, even the difference $x_{ij} - \hat{\pi}_{ij}$ is positive for optimal x_{ij}, the value will be rather small in magnitude. This feature is important for at least two reasons. First, in the small samples, if the value of $\hat{\pi}_{ij}$ is very low, then intuitively, the optimal value of x_{ij} also tends to be very low. Second, the "zero problem" is usually addressed by adding very low numbers instead of zeros. The intuition above suggests that the optimal value x_{ij} of this event will be close to zero.

In the case of the small dataset, an individual with rare characteristics will most likely be underrepresented. The analysis above implies that the IPF algorithm will discriminate against this type of individual, and in the synthetic dataset, an individual will be even more underrepresented. For some applications, it can be an unpleasant issue. To address the problems, we suggest changing the objective function to Beta divergence. Beta divergence (Févotte and Idier (2011)) is defined as

$$D_\beta = \begin{cases} \frac{1}{\beta(1-\beta)} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(x_{ij}^{\beta} + (\beta - 1)\hat{\pi}_{ij}^{\beta} - \beta x_{ij} \hat{\pi}_{ij}^{\beta-1} \log \frac{x_{ij}}{\hat{\pi}_{ij}} \right), & \text{if } \beta \in \mathbb{R} \setminus \{0, 1\}, \\ \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij} \log \frac{x_{ij}}{\hat{\pi}_{ij}}, & \text{if } \beta = 1, \\ \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{x_{ij}}{\hat{\pi}_{ij}} - \log \frac{x_{ij}}{\hat{\pi}_{ij}} - 1 \right), & \text{if } \beta = 0. \end{cases}$$

Standardly, the value of D_β in $\beta = 0$ and $\beta = 1$ is defined by continuity. If $\beta = 1$, then Beta divergence is simply KL divergence.

We observe that for a general value of β, Beta divergence mitigates the problems we discussed in Section 2.2. There is no infinite penalty when $\hat{\pi}_{ij} = 0$ and a large penalty for large value x_{ij} when $\hat{\pi}_{ij}$ is very low. Moreover, let $\hat{\pi}_{ij}$ be zero or very low. The value of β controls the penalty for increasing x_{ij}. The first-order penalty of a pair $(x_{ij}, \hat{\pi}_{ij})$ is controlled by the value $\beta x_{ij}^{\beta-1}$. For larger values of β, where $\beta > 1$, the penalty for larger x_{ij} becomes less severe. Therefore, the analyst can straightforwardly control the penalties. If, in the particular application, the individuals with rare characteristics are less important, the low value of β is more appropriate. If the individuals with rare characteristics are more valuable, the analyst should increase the value of β. The choice of the value of β reminiscences the choice of the hyperparameters in Machine Learning.

We confirm the given above intuition with a toy example.

3 Toy Example

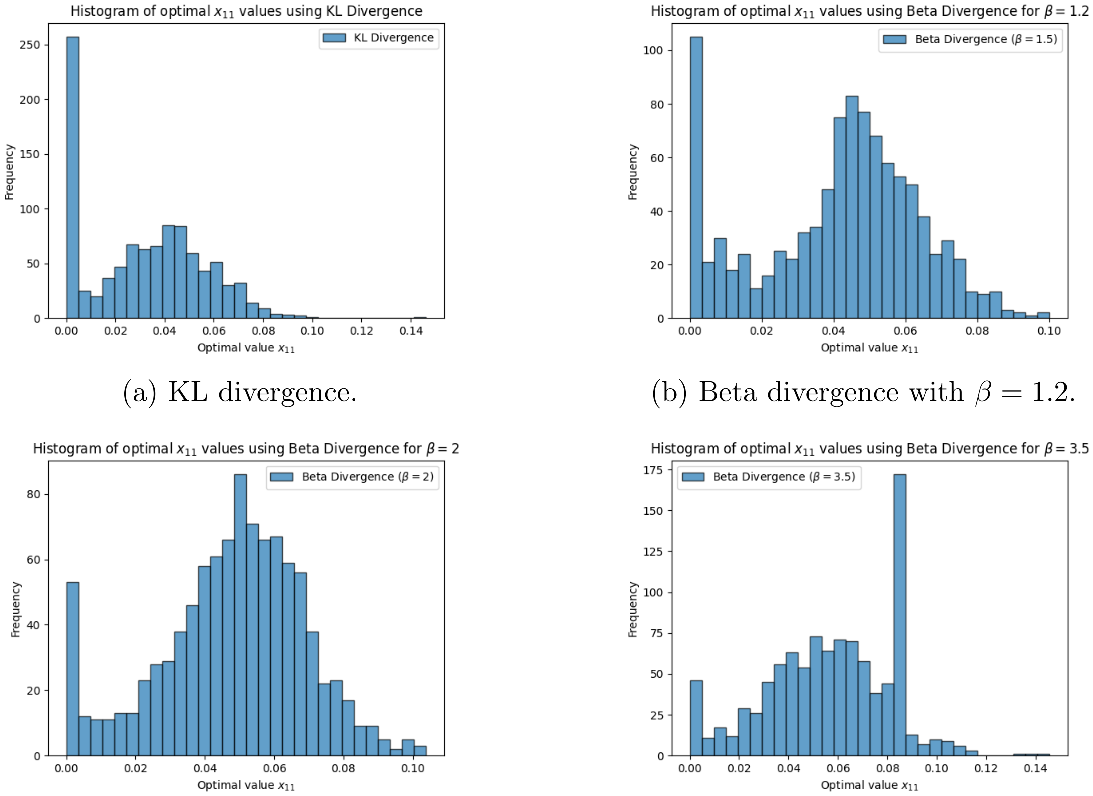
Let (π_{ij}) be the true joint distribution of the population with two characteristics: age and gender. (π_{ij}) is given by matrix

Age/Gender	**Woman**	**Man**	p
Young	0.028	0.139	0.167
Adult	0.25	0.25	0.25
Old	0.222	0.111	0.333
q	0.5	0.5	

Table 1: Matrix representing the distribution (π_{ij}) of age and gender.

We pick the following distribution to analyze the prediction of the probability of an individual with characteristics: Young, Woman. We assume that there is a small sample of size 50. The probability of being in the sample for this type of individual is very low, and it is very likely that such an individual does not even appear in the sample. Therefore, distribution (π_{ij}) is suitable for illustrating the advantages of using Beta divergences.

For our simulations we generate 1000 samples with size 50. For each sample, we calculate the optimal value of x_{11} using four different objective functions: KL divergence and Beta divergence for $\beta = \{1.2, 2, 3.5\}$. We draw a histogram for optimal values of x_{11} for each optimization problem.



(a) KL divergence.

(b) Beta divergence with $\beta = 1.2$.

(c) Beta divergence with $\beta = 2$.

(d) Beta divergence with $\beta = 3.5$.

Figure 1: Histograms for optimal (π_{ij}) for different divergences.

In Figure 1, we observe that, indeed, the optimal (π_{ij}) in the case of KL divergence has a hike peak for the very low values. Other histograms are more spread. Notably, going from value $\beta = 1$, which corresponds to the KL divergence, to $\beta = 1.2$, decreases the amount of very low values by more than half.

Additionally, we observe that the naive increase value of β decreases the small values of (π_{ij}); the distribution does not reflect the true value of the parameter. Therefore, the choice of appropriate β needs to be cautious.

Comparing the fit of the four different models is not a straightforward task because the Beta divergence with $\beta = 2$ is essentially an Euclidean distance and, therefore, MSE as a measure of fit is not appropriate. Instead, we choose the average KL divergence between the fitted distribution and the true one. The results for the comparison are in the Table 2.

KL	**Beta $\boldsymbol{\beta = 1.2}$**	**Beta $\boldsymbol{\beta = 2}$**	**Beta $\boldsymbol{\beta = 3.5}$**
0.0387	0.0236	0.0257	0.0412

Table 2: Comparison of KL and Beta divergences for different values of β.

The results are rather surprising. The minimization of KL divergence is worse than the minimization of Beta divergence with $\beta = \{1.2, 2\}$ in the case of the average KL divergence! The results confirm our intuition that in the case of the small samples appropriately chosen, Beta divergence outperforms KL divergence as a measure.

Finally, we briefly comment that the solution algorithm to problem (1) with Beta divergence can be easily extended from the IPF. Dhillon and Tropp (2008) show how the algorithm for minimization of the Bregman divergences with respect to the affine constraint naturally extends IPF.

References

Bresch, J. and Stein, V. (2024). Interpolating between optimal transport and kl regularized optimal transport using r\'enyi divergences. *arXiv preprint arXiv:2404.18834*.

Chang, S., Qu, Z., Leskovec, J., and Ugander, J. (2023). Inferring networks from marginals using iterative proportional fitting. In *The Second Learning on Graphs Conference*.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Dhillon, I. S. and Tropp, J. A. (2008). Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146.

Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456.

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55(1):179–188.

Kaddoura, I., Masson, D., Hettinger, T., and Unterfinger, M. (2024). An agent-based simulation approach to investigate the shift of switzerland's inland freight transport from road to rail. *Transportation*, 51(5):1701–1722.

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, 14(1):22–59.

Lederrey, G., Hillel, T., and Bierlaire, M. (2022). Datgan: Integrating expert knowledge into deep learning for synthetic tabular data. *arXiv preprint arXiv:2203.03489*.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. (2017). Tsallis regularized optimal transport and ecological inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.