

MIAMI: MIxed Data Augmentation MIxture

Robin Fuchs[1(✉)], Denys Pommeret[1,2], and Samuel Stocksieker[2]

[1] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
{robin.fuchs,denys.pommeret}@univ-amu.fr
[2] Lyon 1 Univ, ISFA, Lab. SAF EA2429, 69366 Lyon, France
samuel.stocksieker@univ-amu.fr

Abstract. Performing data augmentation for mixed datasets remains an open challenge. We propose an adaptation of the Mixed Deep Gaussian Mixture Models (MDGMM) to generate such complex data. The MDGMM explicitly handles the different data types and learns a continuous latent representation of the data that captures their dependence structure and can be exploited to conduct data augmentation. We test the ability of our method to simulate crossings of variables that were rarely observed or unobserved during training. The performances are compared with recent competitors relying on Generative Adversarial Networks, Random Forest, Classification And Regression Trees, or Bayesian networks on the UCI Adult dataset.

Keywords: Mixed data · Data augmentation · Mixture models · Unbalanced data

1 Introduction

Data augmentation is a powerful methodology to deal with unbalanced data, with data containing missing values, as well as to produce synthetic and anonymous datasets. Most data augmentation approaches are designed for a single data type: either continuous or non-continuous, with a particular focus on the continuous case. In the continuous data framework, the main methods are k-nearest neighbors (kNN), perturbation methods adding random noises to the data [13,16], methods based on the dependence structure obtained by modeling joint distribution or copulas [25], Gaussian Mixture Models [24], Generative Adversarial Networks (GAN) [14,22], and Variational Autoencoder (VAE) [17]. For non-continuous data, methods often rely on kNN [7] using adapted metrics, Classification And Regression Trees (CART) or Random Forest [21].

Methods dealing with each data type separately aim at capturing the dependence structure of the observations and using it to generate data. However,

Granted by the Research Chair DIALog under the aegis of the Risk Foundation, an initiative by CNP Assurances.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13375, pp. 113–129, 2022.
https://doi.org/10.1007/978-3-031-10522-7_9

when the data are mixed, performing data augmentation can be challenging since the approaches have to simultaneously model categorical, binary, ordinal, discrete, and continuous data. These mixed distributions often contain multimodal marginal densities, non-standard continuous distributions, and unbalanced modalities for the binary, categorical and ordinal variables.

There exist several recent works in the literature dedicated to the problem of mixed data. Some of them are adaptations of the previously cited methods, such as kNN with a specific distance [3], probabilistic models based on conditional copulas as synthetic data generators [10], or conditional GANs [4,27]. This generalization of GANs was introduced to overcome the fact that traditional GANs had difficulties reproducing complex distributions such as multimodal distributions (e.g. mixture distributions) or modeling entire distributions and to keep the full dependence structure of the data [15]. The Bayesian framework also constitutes a powerful family of methods to deal with mixed data. Dirichlet process mixtures can be used as latent spaces to generate data [9,20]. The Bayesian framework can also be combined with Gaussian copulas to generate fully-synthetic mixed data [5]. Yet, one of the main difficulties of Bayesian models remains in the choice of the priors to reflect the underlying model and the complexity of the dependence structure.

In this work, we introduce an approach based on the Mixed Deep Gaussian Mixture Model (MDGMM) [6]. The MDGMM learns a continuous representation of the dataset and can be inverted to generate pseudo-observations. The proposed methodology keeps the dependence structure of mixed datasets in a flexible way considering the flexible parametric distribution of the latent space which relies on Deep Gaussian Mixture Models [26]. Furthermore, all mixed data types are handled explicitly, especially the ordinal data type that is often assimilated to categorical or continuous data by competitor methods. The MDGMM is hence used as a data generator and coupled with an acceptation-rejection procedure to select observations presenting the desired characteristics. Our main objective is here to reconstruct unobserved regions of the mixed multivariate support of the data. We call this complete procedure "MIAMI", standing for "MIxed data Augmentation MIxture".

2 MDGMM Brief Presentation

The MDGMM is an unsupervised multi-layer model designed for mixed data clustering introduced by [6]. The mixed data are mapped into a continuous latent space using a Generalized Linear Latent Variable Model (GLLVM) [2,18,19]. The latent space is a Deep Gaussian Mixture Model [26] which enables the latent space to capture a broad range of possible distributions. More formally, denoting $Y = (Y_1, \cdots, Y_n)$ the n observations of dimension p, we have for $i \in [1, n]$:

$$\begin{cases} Y_i \to z_i^{(1)} \text{ through GLLVM link via } (\lambda^{(0)}, \Lambda^{(0)}) \\ z_i^{(1)} = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(2)} + u_{i,k_1}^{(1)} \text{ with probability } \pi_{k_1}^{(1)} \\ \ldots \\ z_i^{(L-1)} = \eta_{k_{L-1}}^{(L-1)} + \Lambda_{k_{L-1}}^{(L-1)} z_i^{(L)} + u_{k_{L-1}}^{(L-1)} \text{ with} \\ \text{probability } \pi_{k_L}^{(L-1)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases} \tag{1}$$

where "GLLVM link" refers to the link functions relating the original mixed variable space to the continuous latent space. These link functions $f(Y_i|z_i^{(1)}, \Theta)$ are part of exponential families and the parameters Θ are learned during training (more details are given in [6]). To illustrate the possible link functions, if the jth component of the ith observation, Y_{ij}, is a count variable, one can choose a Binomial distribution:

$$f(Y_{ij}|z^{(1)}, \Theta) = \binom{n_j}{Y_{ij}} f(z^{(1)})^{Y_{ij}} (1 - f(z^{(1)}))^{n_j - Y_{ij}}, \tag{2}$$

with n_j the upper bound of the count variable support. Other examples of link functions are given in [2].

In the simulations, the ordinal variables are linked to the latent space using ordered multinomial distributions, the categorical variables using unordered multinomial distributions, the count variables using Binomial distributions, the binary variables with Bernoulli distributions, and the continuous variables with Gaussian distributions.

The graphical model of the MDGMM described in (1) is presented in Fig. 1. This architecture was introduced as the M1DGMM in [6]. There exists a second architecture, M2DGMM, which merges the embeddings learned separately on continuous and non-continuous variables. We have chosen a simple one-layer deep M1DGMM architecture with a two-dimensional latent space and $K_1 = 4$ components, which has proven to be the more stable architecture [6] and will ensure to obtain more reproducible results.

3 Data Augmentation Procedure

The latent representation of the data is first determined by training the MDGMM on the data Y. The parameters learned are hereafter denoted by a tilde. The model is then inverted to generate pseudo-observations and only the observations with the desired characteristics are kept.

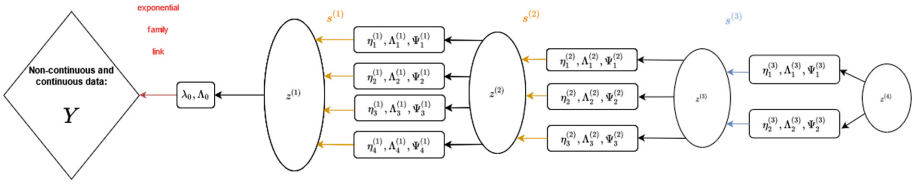


Fig. 1. Graphical model of a M1DGMM

Model training is described in the original MDGMM and enables drawing $z^{(1)}$ samples from a $DGMM(\tilde{\Theta})$ distribution. The model is then inverted using Bayes rule:

$$f(Y|\tilde{\Theta}) = \frac{f(\tilde{z}^{(1)}|\tilde{\Theta}) f(Y|\tilde{z}^{(1)}, \tilde{\Theta})}{f(\tilde{z}^{(1)}|Y, \tilde{\Theta})} \tag{3}$$

$$\propto f(\tilde{z}^{(1)}|\tilde{\Theta}) \prod_{j=1}^{p} f(Y_j|\tilde{z}^{(1)}, \tilde{\Theta}). \tag{4}$$

The passage from (3) to (4) comes from the fact that, by construction, there is mutual independence between the original variables given the latent variable. This means that the latent representation captures all the dependence structure existing in the original dataset, which is a suitable feature to perform data augmentation in the mixed data case.

Let C be the set of wanted characteristics, being for example a region of the original variable space, missing crossings of variables, or unbalanced modalities for non-continuous variables. One can simulate the N^* pseudo-observations presenting the characteristics C using the procedure described in Algorithm 1.

Algorithm 1. MIAMI

Input: Y, N^*, C, m_0, m_1.
Initialize $s = 0$.
repeat
  Generate m_0 draws of $z^{(1)}$ from $DGMM(\tilde{\Theta})$.
  Use these draws to sample m_1 pseudo-observations from $f(Y^*|\tilde{z}^{(1)}, \tilde{\Theta})$ using (4).
  for $i = 1$ **to** m_1 **do**
    if Y_i^* satisfies condition C **then**
      Add Y_i^* to Y^*.
      $s = s + 1$.
    end if
  end for
until $s \geq N^*$
Output: N^* draws of Y^*.

In this algorithm, "Y^* satisfies condition C" means that the pseudo-observations (Y^*) present the wanted characteristics (C). In this sense, Algorithm 1 can be viewed as an oversampling method creating pseudo-observations with the desired features. The N^* pseudo-observations can be simulated by changing the number of copies of the latent variables $z^{(1)}$, m_0, or the number of pseudo-observations Y^*, m_1, to draw from each $z^{(1)}$. More $z^{(1)}$ draws ensure better coverage of the latent space while more Y^* draws per $z^{(1)}$ give more information about the link existing between each latent point and the original variable space.

4 Numerical Illustration

4.1 Competitors

We propose to compare our approach with four recent competitors:

- *CTGAN* [1] is part of the SDV project and relies on a GAN-based Deep Learning data synthesizer to deal with continuous as well as categorical data.
- *Synthpop* proposed in the synthpop package in R [21]. Operating in a non-parametric framework, Synthpop generates the synthetic dataset sequentially by using either a CART procedure or a Random Forest (RF) approach. It is suitable for continuous as well as categorical, ordinal, and binary data. We consider both approaches, namely SynthPop-CART and SynthPop-RF, as competitors.
- *DataSynthesizer* [23] in the DataSynthesizer package in Python captures the underlying correlation structure between the different attributes through a Bayesian network and then draws samples from this model. It is suitable for continuous as well as categorical or binary data.

It is worth pointing out that we had also tested extensions of the SMOTE algorithm [3]: the SMOTE-NC algorithm with the HEOM distance and the so-called Adasyn algorithm [8]. However, these methods obtained much worse results than the other competitors on this dataset. Thus, their results are not shown here.

4.2 Evaluation Metrics

The model performances are here evaluated graphically and by using properly defined metrics. First, the dependence structure between couples of variables can be graphically assessed using Associations Matrices (AM) which are a generalization of correlations matrices for mixed data. In AM, the standard correlations are used to compare pairs of ordered variables (continuous, ordinal, and count variables), correlation ratios are used to compare an ordered variable with a non-ordered variable, and the Cramer's V to compare two non-ordered variables.

Secondly, the performances are also measured by three metrics:

- The association distance which is the Mean Relative Absolute Errors (MRAE) between the test and the generated datasets obtained by summing the absolute relative differences between the values of their association matrices.
- The MAE (Mean Absolute Error) between proportions for binary and categorical variables.
- The Kullback-Liebler divergence between the multivariate continuous distributions of the test and generated datasets.

The association distance hence summarizes how well each method captures the dependence structure, the MAE the quality of the marginal distributions reconstruction for categorical and binary variables, and the Kullback-Liebler a pseudo-distance between the multivariate continuous distributions.
The presented results are obtained over ten runs for each competitor and the formulas of the MAE and Kullback-Liebler divergence are given in Appendix.

4.3 Dataset

We test our approach on the Adult Census Income data. This dataset contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau [11]. The dataset contains n = 32.561 observations and is composed of three continuous variables, six categorical variables, two binary variables, and three ordinal variables. A detailed list of the variables is given in Appendix according to the UCI documentation.

4.4 Experimental Designs

In this work, the ability of the competitor models to generate observations presenting a given combination of two and three variables is tested. This combined modality is either weakly present (10 observations, called "Unbalanced design" hereafter) or completely missing (hereafter "Absent design") in the training set. We have then four designs:

- Absent design for a bivariate modality ("Bivariate Absent"),
- Absent design for a trivariate modality ("Trivariate Absent"),
- Unbalanced design for a bivariate modality ("Bivariate Unbalanced"),
- Unbalanced design for a trivariate modality ("Trivariate Unbalanced").

The bivariate modality is in our case, women of more than 60 years old (age>60 & sex=="Female") and the trivariate modality is widowed women of more than 60 years old (age>60 & sex=="Female" & Marital.status=="Widowed").

The stability of the methods is evaluated using a 10-fold approach: for each experimental design, ten training sets of 1000 observations are drawn from the original dataset. The test sets are composed of the observations presenting the desired modality and that are not included in the train set. The number of pseudo-observations presenting the desired crossing of variables N^* to draw is 200 for each competitor.

4.5 Results Analysis

We restrict our attention here to the most representative results.

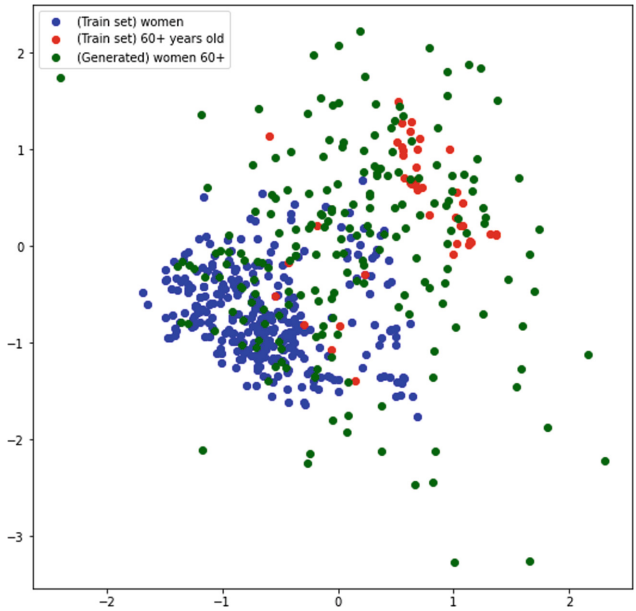


Fig. 2. Latent representation of women (in blue) and 60+ years old individuals (in red) coming from the train dataset, and women of 60+ years old generated by MIAMI (in green) in the Absent bivariate design. (Color figure online)

As shown in Fig. 2, the characteristics of the original dataset are well mapped into the latent space ($z^{(1)}$), which is the first layer of the MDGMM as illustrated in Fig. 1. Women and individuals of more than 60 years old are represented in two different and coherent zones. The generated individuals which present both characteristics are mainly generated near these two zones denoting that this global area well encodes the modality crossing.

Concerning the reproduction of the dependence structure through the association distances (Fig. 3), CTGAN generally obtains the best performance followed by MIAMI, DataSynthesizer, and SynthPop-CART. MIAMI is especially competitive on the Absent bivariate and Unbalanced trivariate designs. The SynthPop-RF approach outperforms most methods on trivariate designs but fails to capture the dependence structure of the bivariate designs.

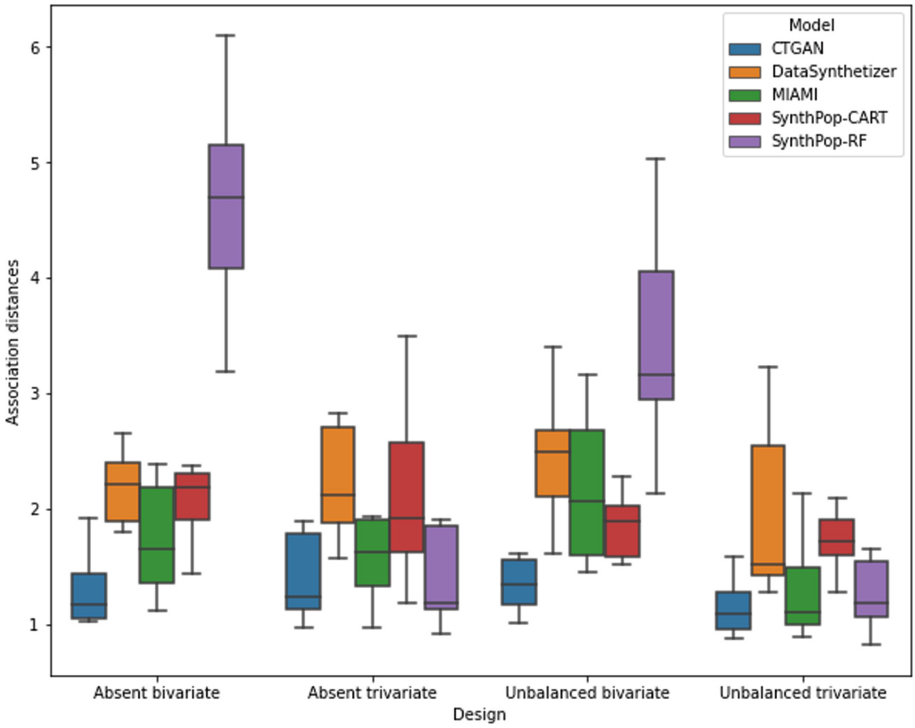
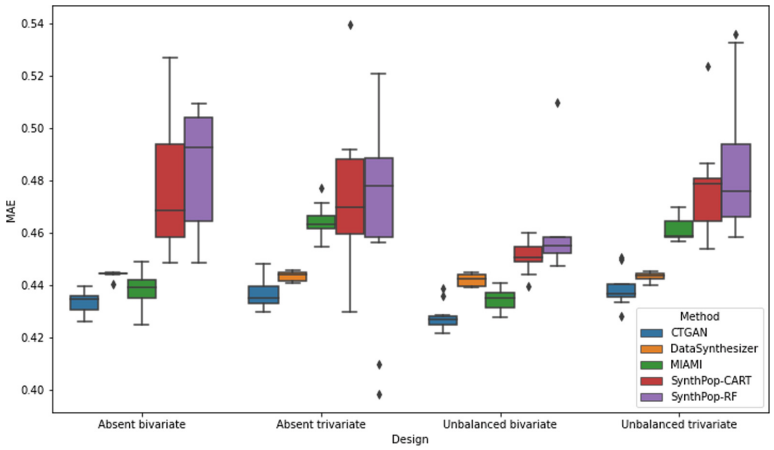


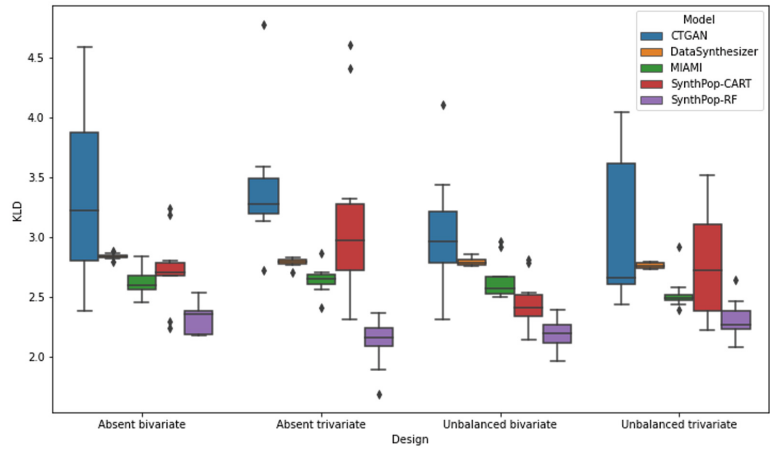
Fig. 3. Association distance for the four designs. The lower the distance, the best the dependence structure is reproduced.

CTGAN and MIAMI still obtain the best results for the reconstruction of categorical and binary variables but DataSynthesizer also performs well (Fig. 4(a)). The SynthPop methods present a much higher variance than the other competitors for three designs out of four, while MIAMI and CTGAN obtain comparable variances and DataSynthesizer a much lower variance.

Besides, the continuous variables are best accounted for by SynthPop-RF, MIAMI, and SynthPop-CART (Fig. 4(b)). The distributions generated by CTGAN are furthest from the test ones and show more variability, especially in comparison with DataSynthesizer and MIAMI. This pattern can also be observed in Fig. 5 representing the bivariate distribution of the age and flnwgt variables on the test dataset and for the observations generated by MIAMI and CTGAN. The continuous distributions reconstructed by MIAMI are more concentrated and the density maximum is closest to the one of the test dataset when compared to CTGAN.



(a) MAE



(b) Multivariate Kullback-Liebler

Fig. 4. MAE (a) and Kullback-Liebler divergence (b) between the test and the generated datasets for all designs

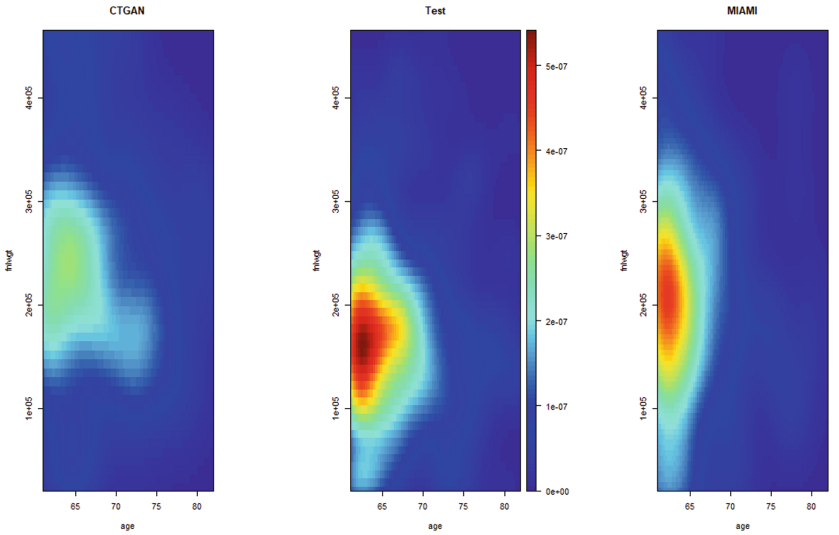


Fig. 5. Density estimation of the bivariate distribution (Fnlweight, Age) for CTGAN, on the test set, and for MIAMI for the Unbalanced bivariate design.

Figure 6 gives an illustration of the generation of the seven modalities of the variable "Marital Status" for the Bivariate Unbalanced and Absent designs. It can be seen that only MIAMI and DataSynthesizer manage to generate all the possible modalities in the first design. MIAMI is closest to the observations while DataSynthesizer creates unobserved modalities. CTGAN is concentrated on only one modality. In the Absent design, none of the methods manage to cover all the modalities while DataSynthesizer once again proposes an unobserved modality.

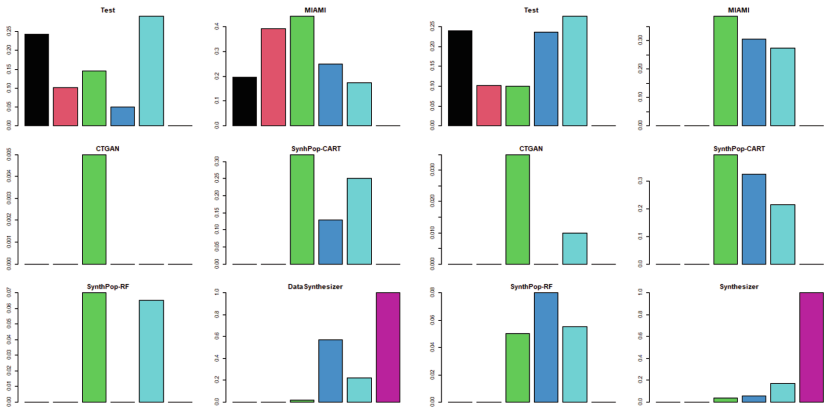


Fig. 6. Modality proportions for the Marital-status under the Bivariate Unbalanced design (left) and the Bivariate Absent design (right)

Finally, compared to the competitor methods, MIAMI gives associations close to the true associations existing in the test set but with a slightly lower intensity than in the test set (Fig. 7). The associations between "Education.num" and "Occupation" or the one between the marital status and the "Relationship" variable are captured. These two associations are also well reproduced by SynthPop-CART, which tends however to create nonexistent associations between most variables. CTGAN has more difficulty in reproducing the original patterns. For concision purposes, the results of SynthPop-RF and DataSynthesizer are not presented. Indeed, SynthPop-RF exaggerates the associations existing between variables even more than SynthPop-CART, and DataSynthesizer fails to reproduce the main patterns of the test association matrix.

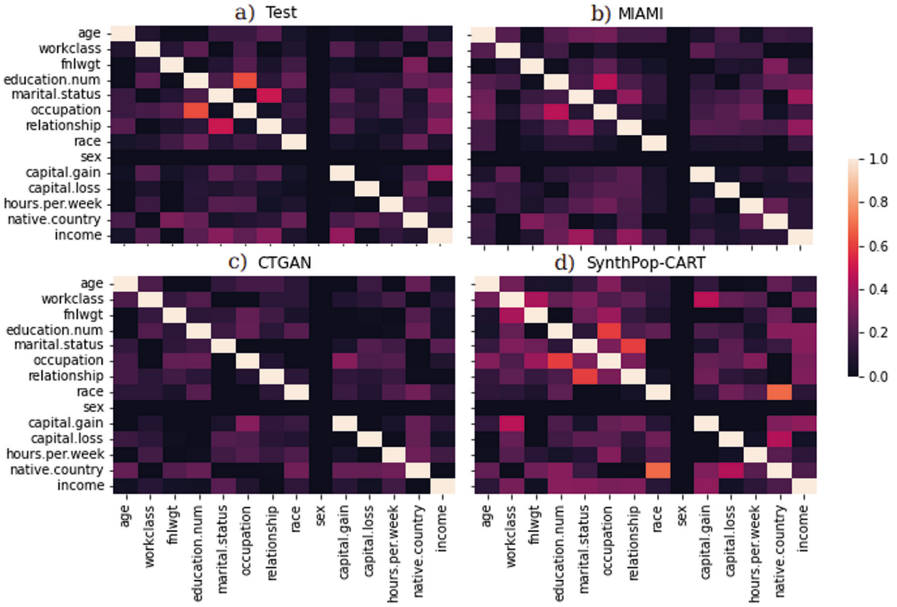


Fig. 7. Association matrices for women of more than 60 years old existing in the test set of the Unbalanced bivariate design (a), generated by MIAMI (b), by CTGAN (c) and by SynthPop-CART (d).

The code to reproduce the results is available at https://github.com/RobeeF/M1DGMM.

5 Discussion and Perspective

MIAMI is an algorithm dedicated to mixed data which oversamples desired areas of the sampling space while preserving the multivariate dependence structure of the data. Based on our numerical study, we conclude that MIAMI seems

to reconstruct well the joint dependence of the mixed data, the univariate non-continuous distributions, and the multivariate continuous distributions. Its major competitor seems to be CTGAN in terms of dependence structure and non-continuous variable, and SynthPop-CART on the multivariate continuous distributions reconstruction.

The flexibility of its latent space enables MIAMI to properly reconstruct areas of high density for continuous variables and to generate a wide range of modalities for non-continuous variables while the other methods often reproduce only the most represented ones (Fig. 6 and Figs. 8 and 9 in Appendix). In this work, the latent space takes the form of a simple one hidden layer architecture. More complex architectures and hyperparameters could be investigated in future works.

Furthermore, MIAMI here generated pseudo-observations using the entire latent space, and only the pseudo-observations presenting the desired characteristics were kept. However, as shown in Fig. 2, some regions of the latent space are more likely than others to generate these pseudo-observations of interest. Hence, the sampling of the latent space could be adapted through the procedure to increase the generation/acceptation ratio. One could for instance exploit the link functions of the MDGMM or rely on Bayesian optimization methods. In the latter solution, the distribution of the latent variable would be taken as a prior and the task will be to estimate the posterior areas presenting high-acceptation rates.

Appendix

A Datasets Details

The variables of the Adult dataset are according to the UCI documentation:

- Income: binary (>50K, <=50K).
- Age: continuous.
- Workclass: categorical (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked).
- Fnlwgt: continuous.
- Education-num: ordinal.
- Marital-status: categorical (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse).
- Occupation: categorical (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces).
- Relationship: categorical (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).
- Race: categorical (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).

- Sex: binary (Female, Male).
- Capital-gain: ordinal.
- Capital-loss: ordinal.
- Hours-per-week: continuous.
- Native-country: categorical (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc.), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands).

B Evaluation Metrics Details

Our overall criterion between a test dataset and a generated dataset is the association distance obtained as follows:

$$DA = \frac{1}{P} \sum_{1 \leq i < j \leq p}^{p} |M_{ij}(test) - M_{ij}(gen)| / M_{ij}(test),$$

where p denotes the number of variables (14 in the Adult dataset), $P = (p^2 - p)/2$, and $M_{ij}(test)$ (resp. $M_{ij}(gen)$) is the (i, j)th entry of the test (resp. generated) Association Matrix.

To measure the similarity between the dependence structures of the vectors formed by the three continuous variables (Age, Fnlwgt, and Hours) we used the multivariate Kullback Leibler divergence [12].

For qualitative data, we chose the mean absolute errors (MAE) between proportions. More precisely, for a kth intersection of modalities we consider

$$MAE(k) = |p_k(test) - p_k(gen)|,$$

where $p_k(test)$ (resp. $p_k(gen)$) stands for kth test (resp. generated) proportion. The final MAE is the mean of all the $MAE(k)$ over all the possibilities.

C Additional Results: Unobserved Marginal Density Reconstruction

When it comes to reconstructing univariate densities, MIAMI generates well-identified unimodal densities contrary to DataSynthesizer which generates flat densities, or SynthPop-CART which generates multi-modal densities (Fig. 8). More precisely, Figs. 8 and 9 represent the estimations of the observed density versus the generated one for Age in the case of the Bivariate and Trivariate Unbalanced designs. We observe that MIAMI can recover the right distribution, yet not observed in the training set. It means that MIAMI captures well the dependence structure of such a partially unobserved variable. CTGAN and

SynthPop-RF also seem to work well for both designs. DataSynthesizer shows a larger variance. This illustration shows that we cannot clearly decide between the methods by looking only at the marginal distributions. Only a criterion like the association distance can take into account a more complex multivariate dependence structure.

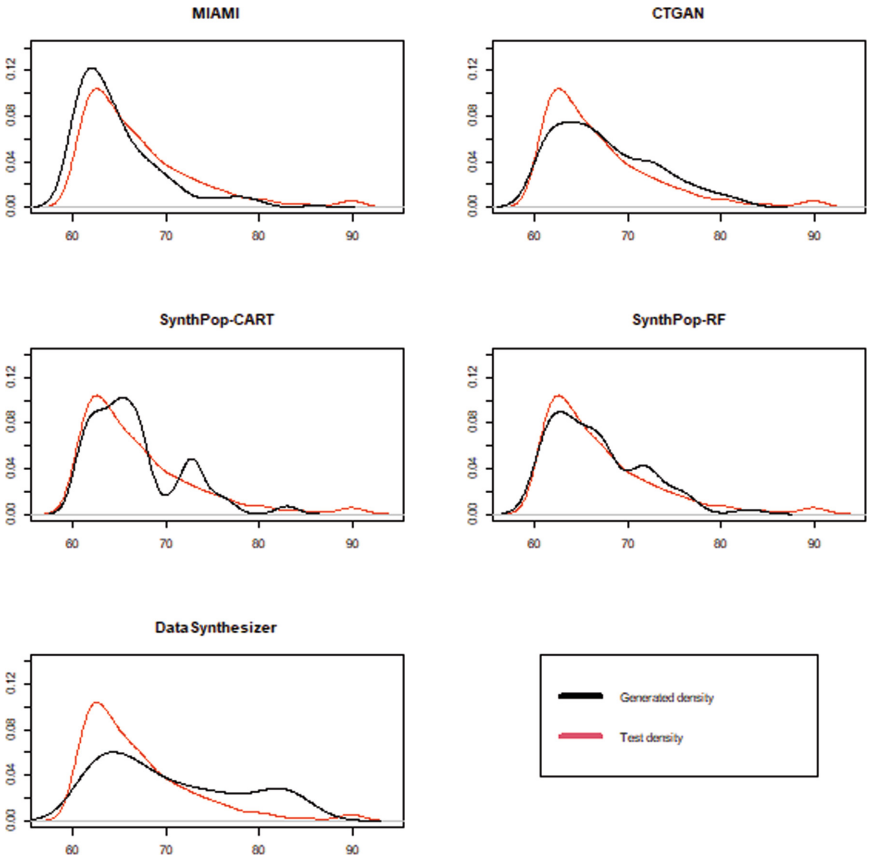


Fig. 8. Density estimations for Age based on the test dataset (red) or based on the Bivariate Unbalanced design (black) (Color figure online)

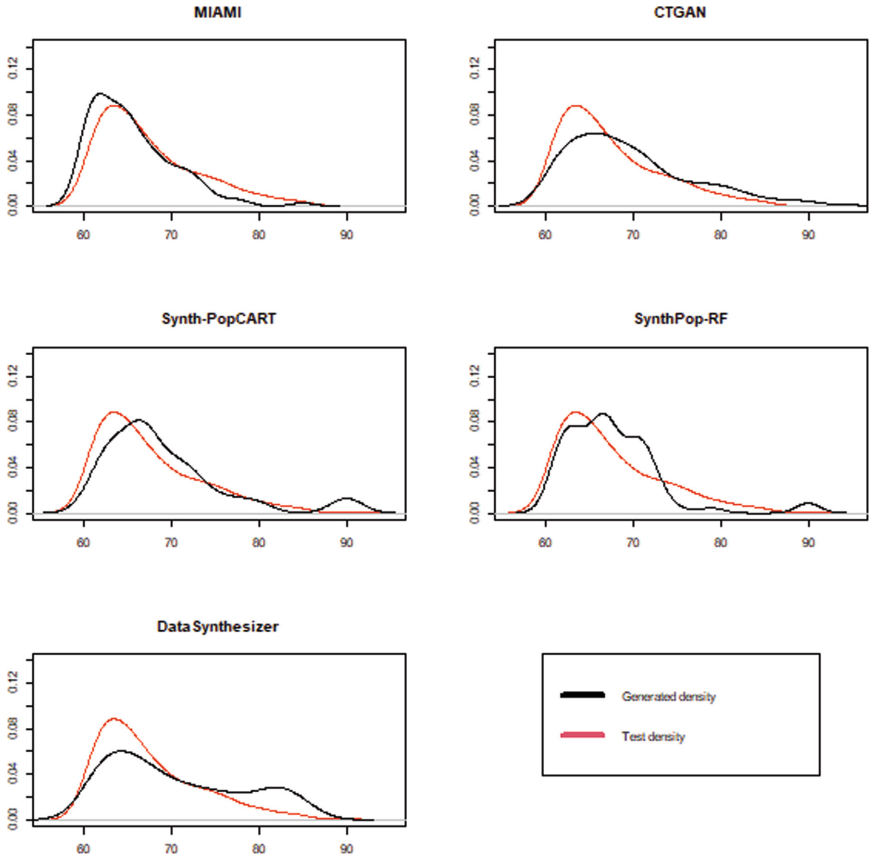


Fig. 9. Density estimations for Age based on the test dataset (red) or based on the Trivariate Unbalanced design (black) (Color figure online)

References

1. Buuren, S.V., Brand, J.P., Groothuis-Oudshoorn, C.G., Rubin, D.B.: Fully conditional specification in multivariate imputation. J. Stat. Comput. Simul. **76**(12), 1049–1064 (2006)
2. Cagnone, S., Viroli, C.: A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. AStA Advances in Statistical Analysis **98**(1), 1–20 (2013). https://doi.org/10.1007/s10182-012-0206-5
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (2002)
4. Engelmann, J., Lessmann, S.: Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. Expert Syst. Appl. **174**, 114582 (2021)
5. Feldman, J., Kowal, D.: A bayesian framework for generation of fully synthetic mixed datasets (2021)
6. Fuchs, R., Pommeret, D., Viroli, C.: Mixed deep gaussian mixture model: a clustering model for mixed datasets. In: Advances in Data Analysis and Classification, pp. 1–23 (2021)

7. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3929–3938. PMLR, 13–18 Jul 2020
8. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IJCNN 2008, pp. 1322–1328 (2008)
9. Hu, J., Reiter, J.P., Wang, Q., et al.: Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. Bayesian Anal. **13**(1), 183–200 (2018)
10. Kamthe, S., Assefa, S., Deisenroth, M.: Copula flows for synthetic data generation (2021)
11. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD 1996, pp. 202–207. AAAI Press (1996)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. **22**(1), 79–86 (1951)
13. Lee, S.S.: Noisy replication in skewed binary classification. Comput. Stat. Data Anal. **34**(2), 165–191 (2000)
14. Liu, Y., et al.: Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. Engineering (2019)
15. Lucic, M., Kurach, K., Michalski, M., Bousquet, O., Gelly, S.: Are gans created equal? a large-scale study. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 698–707. NIPS 2018, Curran Associates Inc., Red Hook (2018)
16. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. **28**(1), 92–122 (2012). https://doi.org/10.1007/s10618-012-0295-5
17. Moreno-Barea, F.J., Jerez, J.M., Franco, L.: Improving classification accuracy using data augmentation on small data sets. Expert Syst. Appl. **161**, 113696 (2020)
18. Moustaki, I.: A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. Br. J. Math. Stat. Psychol. **56**(2), 337–357 (2003)
19. Moustaki, I., Knott, M.: Generalized latent trait models. Psychometrika **65**(3), 391–411 (2000)
20. Murray, J.S., Reiter, J.P.: Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. J. Am. Stat. Assoc. **111**(516), 1466–1479 (2016)
21. Nowok, B., Raab, G.M., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. J. Stat. Soft. **74**(11), 1–26 (2016). https://doi.org/10.18637/jss.v074.i11
22. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. Proc. VLDB Endow. **11**(10), 1071–1083 (2018)
23. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM 2017, Association for Computing Machinery, New York (2017)

24. Richardson, E., Weiss, Y.: On gans and gmms. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018)
25. Sun, Y., Cuesta-Infante, A., Veeramachaneni, K.: Learning vine copula models for synthetic data generation. In: AAAI (2019)
26. Viroli, C., McLachlan, G.J.: Deep gaussian mixture models. Stat. Comput. **29**(1), 43–51 (2019)
27. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. In: NeurIPS (2019)