A Generalized Theory of Mixup
for Structure-Preserving Synthetic Data

Chungpa Lee
Yonsei University

Jongho Im
Yonsei University

Joseph H.T. Kim
Yonsei University

Abstract

Mixup is a widely adopted data augmentation technique known for enhancing the generalization of machine learning models by interpolating between data points. Despite its success and popularity, limited attention has been given to understanding the statistical properties of the synthetic data it generates. In this paper, we delve into the theoretical underpinnings of mixup, specifically its effects on the statistical structure of synthesized data. We demonstrate that while mixup improves model performance, it can distort key statistical properties such as variance, potentially leading to unintended consequences in data synthesis. To address this, we propose a novel mixup method that incorporates a generalized and flexible weighting scheme, better preserving the original data's structure. Through theoretical developments, we provide conditions under which our proposed method maintains the (co)variance and distributional properties of the original dataset. Numerical experiments confirm that the new approach not only preserves the statistical characteristics of the original data but also sustains model performance across repeated synthesis, alleviating concerns of model collapse identified in previous research.

1 INTRODUCTION

Mixup is a prominent data augmentation method (Zhang et al., 2018) that generates new instances by linearly combining observed instances, applicable across both structured and unstructured datasets. By training models on these interpolated samples, mixup enhances the generalization performance of state-of-the-art neural network architectures (Verma et al., 2019; Yun et al., 2019; Guo, 2020; Sohn et al., 2022; Kim et al., 2021; Baena et al., 2022; Chen et al., 2020; Zhang et al., 2022b; Sun et al., 2024). A similar approach, SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002; He et al., 2008; Bunkhumpornpat et al., 2012; Douzas et al., 2018), also leverages interpolated synthetic instances to enhance model performance particularly for imbalanced or long-tail distributions, showcasing the effectiveness of mixup methods.

In this paper we place special focus on data synthesis, an important constituent of data augmentation. While there is extensive research on how synthetic data generated by mixup can enhance model performance (Carratino et al., 2022; Zhang et al., 2021), less attention has been given to understanding the fundamental properties of the synthesized data itself; see Sec. 2.1. In fact most mixup methods generate linearly interpolated instances by taking a weighted average where the weights are randomly drawn from distributions within the range of $[0, 1]$, such as the beta or the uniform distribution. However, this interpolation process reduces the variance, which inevitably distorts the statistical structure of the original dataset both marginally and jointly. The net effect is a less dispersed dataset with more emphasis on representative instances and suppressing the others. In this regard, mixup-based synthetic datasets achieve better performance in training machine learning models from sacrificing non-representative instances, such as the tail instances, in the dataset. Naturally, understanding the impact of mixup warrants further research.

In a similar line of thought, a recent work in Nature (Shumailov et al., 2024) raises concerns regarding the risks associated with training models using data that has been repeatedly synthesized. This phenomenon, known as *model collapse*, describes a situation where the tails of the original distribution are

lost after repeatedly synthesizing the original dataset, demonstrating that over-reliance on synthetic data can lead to catastrophic defects in model training. As the prevalence of synthetic data from generative models increases, therefore, it is essential to carefully consider the quality and structure of this synthetic data to maintain the benefits of training on such datasets.

To this end, in this paper we formally investigate the theoretical properties of mixup and its impact on the resulting synthetic data. Our findings provide insight into how mixup alters the statistical structure of synthetic data in comparison to the original, thereby explaining why standard mixup, while beneficial for improving machine learning model performance, can lead to unintended effects in data synthesis. To address this, we also propose a new mixup method featuring a more generalized and flexible weighting scheme, allowing the synthetic dataset to better preserve the underlying structure of the original data. Our key contributions are as follows:

- We theoretically derive a set of conditions for a mixup weight distribution that preserves the (co)variance for any pair of continuous variables in a general setting.

- We prove that the mean and variance of any numerical variable in a dataset, conditioned on a categorical variable in the same dataset, can be maintained within a specified error bound, which can be controlled by a function of mixup weights.

- As an additional contribution, we propose a new class of mixup weight distributions that satisfy these theoretical conditions, thereby preserving the original data structure with respect to both mean and variance.

- Our numerical experiments show that the proposed mixup method generates synthetic data that preserves fundamental distributional properties, leading to more accurate statistical inferences. Also, regarding model performance with synthetic data, the proposed mixup method yields results comparable to existing synthesis techniques. Notably, it significantly maintains performance under repeated synthesis, addressing concerns raised by Shumailov et al. (2024).

2 RELATED WORK

2.1 Analysis on Mixup

Several theoretical studies have explored the effects of mixup (Carratino et al., 2022; Zhang et al., 2021,

2022a; Park et al., 2022a). In particular, Zhang et al. (2021) and Carratino et al. (2022) demonstrate how training with mixup-generated data enhances model regularization and generalization from the perspective of empirical risk minimization. Furthermore, Zhang et al. (2021) showed that the coefficients of linear least-squares regression are preserved for any synthetic data generated by mixup methods. This preservation property follows directly from the fact that the key statistic in linear regression is the correlation, as shown in Appendix A.1.

A distinctive aspect of our work is the explicit formulation of conditions that preserve the structure of synthetic data. Unlike prior approaches that require post-generation transformations to maintain statistical properties, our method ensures structure preservation during the data generation process, eliminating the need for additional computational steps.

2.2 Synthetic Tabular Data

In the statistical community, various methods for synthesizing tabular data have been extensively studied (Raghunathan et al., 2003; Nowok et al., 2016; Kim et al., 2014; Si and Reiter, 2013; Murray and Reiter, 2016). Advances in deep neural network-based generative models have further led to the development of techniques such as variational autoencoders (Xu et al., 2019; Ma et al., 2020), generative adversarial networks (Park et al., 2018; Choi et al., 2017; Xu et al., 2019; Zhao et al., 2021, 2024; Baowaly et al., 2019), diffusion models (Kotelnikov et al., 2023; Kim et al., 2022; Lee et al., 2023; Kim et al., 2023; Zhang et al., 2024), and large language models (Borisov et al., 2023; Solatorio and Dupriez, 2023; Zhang et al., 2023; Gulati and Roysdon, 2024) for synthesizing tabular data.

Although some of these generative models demonstrate performance comparable to traditional statistical methods, they require significant time and resources for model training. In contrast, our structure-preserving mixup method can generate high-utility synthetic data without the need for training a model. Additionally, it offers the advantage of allowing explicit control over the degree of preservation of the original data structure.

3 PROPERTIES OF SYNTHETIC DATA GENERATED BY MIXUP

In this section, we analysis the statistical properties of synthetic data generated by the mixup method. Especially, we compare the mean and variance of the synthetic data $\tilde{D}$ against those of the original D. Matching these key statistics is crucial as many models rely
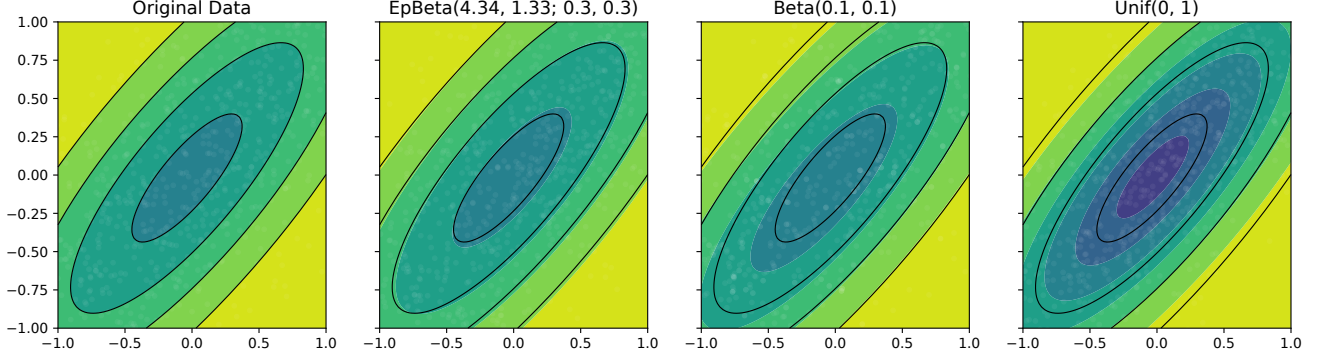
Figure 1: A contour plot of a sample comprising 1000 data points from $N\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right),\left(\begin{smallmatrix}1.1 & 0.9\\0.9 & 1.1\end{smallmatrix}\right)\right)$ is shown in the first plot. This original data is then synthesized with three different mixup weight distributions: the proposed EpBeta$(4.34, 1.33; 0.3, 0.3)$, as well as two standard distributions, Beta$(0.1, 0.1)$ and Unif$(0, 1)$. While the synthetic data generated by Beta or Unif is shrunken, the proposed EpBeta preserves the original data structure.

on assumptions and analyses related to the mean and variance structure; see Sec. 5.

Data typically consists of two types of variables; continuous and categorical. To fix the notation, let us define the original instances as $D_i = (X_i, Y_i, L_i)$ for $i \in [n]$ where $X, Y \in \mathbb{R}$ are continuous variables, and $L \in [c]$ is a categorical variable with c categories. We assume that D_i's are independent and identically distributed according to the population distribution $\mathfrak{D}$. While D_i contains only three variables for our theoretical developments, the results are general and can be easily extended to the cases of multiple continuous and categorical variables.

We denote the synthetic instances generated using the mixup by $\tilde{D}_k = (\tilde{X}_k, \tilde{Y}_k, \tilde{L}_k)$ for $k \in [m]$. To make the source clear, we use indices '$i, j \in [n]$' for each original instance and '$k \in [m]$' for each synthetic instance.

3.1 Synthetic Data Generated by Mixup

Synthetic data using the mixup can be created as follows. First, we write the mixup weights as $W_k^X, W_k^Y, W_k^L \in \mathbb{R}$, each of which is associated to variables X, Y, and L, respectively. These weights are random variables, independent and identically generated from some given distributions $\mathfrak{W}^X, \mathfrak{W}^Y$, and $\mathfrak{W}^L$, respectively. Taking values in $[0, 1]$ is typical for the weights but not required. By construction, these mixup weights are independent of the original instances D_i for all $i \in [n]$. A synthetic instance $\tilde{D}_k = (\tilde{X}_k, \tilde{Y}_k, \tilde{L}_k)$ is obtained by randomly selecting two original instances D_{i_k} and D_{j_k} and applying the following transform:

$$\tilde{X}_k = W_k^X X_{i_k} + (1 - W_k^X) X_{j_k}, \tag{1}$$

$$\tilde{Y}_k = W_k^Y Y_{i_k} + (1 - W_k^Y) Y_{j_k}, \tag{2}$$

$$\tilde{L}_k = \begin{cases} L_{i_k} & \text{if } W_k^L \geq \tau \\ L_{j_k} & \text{if } W_k^L < \tau, \end{cases} \tag{3}$$

where $\tau \in \mathbb{R}$ is the pre-defined cut point for the categorical variable L, with $\tau = 0.5$ being a default choice. Unlike the continuous variables, the mixup for a categorical variable requires a random selection due to its nature. It is noted that common mixup methods impose the same weight, following the original proposal by Zhang et al. (2018). That is, all weights in (1)–(3) are set equal so that

$$W = W_k^X = W_k^Y = W_k^L \text{ for all } k \in [m]. \tag{4}$$

We call (1)–(4) the equal-weight or standard mixup scheme whereas (1)–(3) the general-weight mixup scheme. The standard scheme is a special case of the general-weight scheme, and easier to work with; in fact most theoretical developments in the literature have been based on the standard scheme. We differentiate these two schemes because several theoretical findings in this paper hold under the general-weight scheme, and thus applicable to a more general setting.

In what follows the indices i, j and k are omitted for simple notation provided that there is no confusion.

3.2 Continuous Variable

Under the common mixup technique explained above, it is trivial to prove that the mean of a continuous variable X is always preserved, i.e., $\mathbb{E}[\tilde{X}] = \mathbb{E}[X]$, regardless of the choice of mixup weight distribution:

$$\mathbb{E}[\tilde{X}] = \mathbb{E}[W^X]\mathbb{E}[X] + (1 - \mathbb{E}[W^X])\mathbb{E}[X] = \mathbb{E}[X].$$

For the variance however things are more complicated. In fact we can show that (see Appendix A.2)

$$\text{Var}[\tilde{X}] = \text{Var}[X] + 2\mathbb{E}[(W^X)^2 - W^X]\text{Var}[X],$$

which is generally different from $\text{Var}[X]$. The exact relationship between $\text{Var}[\tilde{X}]$ and $\text{Var}[X]$ is presented below.

Lemma 1 (Variance)**.** *For any synthetic $\tilde{X}$ generated from X in (1):*

1. $\text{Var}[\tilde{X}] = \text{Var}[X]$, *if and only if (iff) the first and second moments of mixup weight are equal. That is*

$$\mathbb{E}\big[(W^X)^2\big] = \mathbb{E}[W^X]. \tag{5}$$

2. $\text{Var}[\tilde{X}] < \text{Var}[X]$, *iff* $\mathbb{E}\big[(W^X)^2\big] < \mathbb{E}[W^X]$.

3. $\text{Var}[\tilde{X}] > \text{Var}[X]$. *iff* $\mathbb{E}\big[(W^X)^2\big] > \mathbb{E}[W^X]$.

All proofs of this paper are provided in Appendix A. Another quantity of interest is the covariance because synthetic data is often required to preserve the correlation of the original data.

Theorem 2 (Covariance)**.** *For any synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the general-weight mixup in (1) and (2):*

1. $\text{Cov}[\tilde{X}, \tilde{Y}] = \text{Cov}[X, Y]$, *iff*

$$\mathbb{E}[W^X W^Y] = \frac{1}{2}\big(\mathbb{E}[W^X] + \mathbb{E}[W^Y]\big). \tag{6}$$

2. $\text{Cov}[\tilde{X}, \tilde{Y}] < \text{Cov}[X, Y]$, *iff* $\mathbb{E}[W^X W^Y] < \frac{1}{2}\big(\mathbb{E}[W^X] + \mathbb{E}[W^Y]\big)$.

3. $\text{Cov}[\tilde{X}, \tilde{Y}] > \text{Cov}[X, Y]$, *iff* $\mathbb{E}[W^X W^Y] > \frac{1}{2}\big(\mathbb{E}[W^X] + \mathbb{E}[W^Y]\big)$.

We comment that Zhang et al. (2021) showed that the coefficients of the linear regression model are preserved under the standard mixup scheme. Though they did not use in their proof, this preservation is essentially a consequence of the correlation-preserving property of the synthetic data when the mixup is conducted with equal weights as in (4); the following result can shorten the proof of Zhang et al. (2021), see Appendix A.1.

Corollary 3 (Correlation)**.** *For any synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the standard mixup scheme, we have*

$$\text{Corr}[\tilde{X}, \tilde{Y}] = \text{Corr}[X, Y]. \tag{7}$$

Before closing this section we emphasize that the results in this section hold generally and equally applicable for data with multiple variables with no further modifications.

3.3 Continuous Variable Conditioned by Categorical Variable

We now bring in the categorical variable L so that we can investigate the synthetic distribution of both continuous and categorical variables jointly. This is motivated by the fact that preserving the mean and variance of the continuous variables conditional on the categorical variable is often necessary in data analyses; *i.e.*, when the height of students in a school exhibit different distributions depending on gender, synthetic height datasets need to preserve their mean and variance for each gender.

Without loss of generality, let us consider a synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup scheme, where X is continuous and L categorical. Our goal is to study the conditional mean $\mathbb{E}[\tilde{X}|\tilde{L}]$. For this, we start with defining a special function of the mixup weights.

Definition 1. *Define the function of general mixup weights W^X and W^L with a cut point τ as*

$$u(W^X, W^L, \tau) = \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\}\big], \tag{8}$$

where $\mathbf{I}$ is an indicator function. Under the standard mixup scheme this reduces to

$$u(W, \tau) = \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq \tau\} + W\mathbf{I}\{W < \tau\}\big]. \tag{9}$$

With this function we can show that the synthetic conditional mean $\mathbb{E}[\tilde{X}|\tilde{L} = l]$ is a convex combination of the original conditional mean $\mathbb{E}[X|L = l]$ and the marginal mean $\mathbb{E}[X]$.

Theorem 4 (Conditional Mean)**.** *For any synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup, where X is continuous and L is categorical, the synthetic conditional mean $\mathbb{E}[\tilde{X}|\tilde{L} = l]$ can be expressed as*

$$\mathbb{E}[\tilde{X}|\tilde{L} = l] = (1 - u(W^X, W^L, \tau)) \cdot \mathbb{E}[X|L = l] + u(W^X, W^L, \tau) \cdot \mathbb{E}[X], \tag{10}$$

or, alternatively,

$$\mathbb{E}[\tilde{X}|\tilde{L} = l] = \big(1 - u(W^X, W^L, \tau)\Pr\{L \neq l\}\big) \\ \cdot \mathbb{E}[X|L = l] \\ + u(W^X, W^L, \tau)\Pr\{L \neq l\} \\ \cdot E[X|L \neq l]. \tag{11}$$

Both expressions in Theorem 4 are weighted sums of two terms with the same first term $\mathbb{E}[X|L = l]$, indicating that $\mathbb{E}[\tilde{X}|\tilde{L} = l]$ partly uses the same information of the original data. The second terms however

are different. In (10) it is given as $\mathbb{E}[X]$ which can be understood as the overall information of X, whereas $\mathbb{E}[X|L \neq l]$ in (11) is seen as complimentary information of the original data. This aligns with a general principle found in the data augmentation literature, which says that generating data leverages additional or overall information of the given data; see Bowles et al. (2018) and Mumuni and Mumuni (2022). Also, focusing on (11), large $\Pr\{L \neq l\}$ value puts more weight on $\mathbb{E}[X|L \neq l]$, suggesting that this probability can measure the credibility of the additional information.

Turning to the accuracy of the conditional mean of the synthetic data relative to the original one, we present the following result.

Corollary 5 (Conditional Mean Gap). *For any synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup, where X is continuous and L is categorical, the difference between the conditional mean is given by*

$$\left|\mathbb{E}[\tilde{X}|\tilde{L}=l] - \mathbb{E}[X|L=l]\right|$$
$$= \left|u(W^X, W^L, \tau)\right| \cdot \Pr\{L \neq l\}$$
$$\cdot \left|\mathbb{E}[X|L=l] - \mathbb{E}[X|L \neq l]\right|. \quad (12)$$

The right side of (12) consists of three terms, where the first term can be controlled by the modeler and the remaining two terms are fixed for given data. Thus an important observation on this result is that the difference in the left side of (12) can be made smaller by controlling the value of $u(W^X, W^L, \tau)$ with suitably chosen weight random variables and the cut point.

Assuming that the variance of the synthetic data is preserved, we can establish the following upper bound for the conditional variance.

Theorem 6 (Conditional Variance Gap). *Assume that $\mathbb{E}[(W^X)^2] = \mathbb{E}[W^X]$. Then, for any synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup, where X is continuous and L is categorical, the difference between the conditional variance is bounded as follows:*

$$\left|\mathrm{Var}[\tilde{X}|\tilde{L}=l] - \mathrm{Var}[X|L=l]\right|$$
$$\leq \left|u(W^X, W^L, \tau)\right| \cdot \left|\mathrm{Var}[X|L=l] - \mathrm{Var}[X]\right|$$
$$+ \left|u(W^X, W^L, \tau)(1 - u(W^X, W^L, \tau))\right|$$
$$\cdot \left(\mathbb{E}[X|L=l] - \mathbb{E}[X]\right)^2. \quad (13)$$

Similar to Corollary 5, one can make the difference of the conditional variance smaller by controlling the right side of (13) where the only quantity at the modeler's disposal is $u(W^X, W^L, \tau)$. In Sec. 4 we propose a class of mixup weight distributions that has a explicit relationship to this function.

Our discussion so far shows that function $u(W^X, W^L, \tau)$ in Def. 1 plays an important role in computing the conditional moments in the synthetic data. Thus we present two theoretical properties of this function before concluding this section.

Lemma 7. *Under the standard mixup scheme, $u(W, \tau) \in [0, 1]$ holds for any $\tau \in \mathbb{R}$ if $\mathbb{E}[W^2] = \mathbb{E}[W]$.*

Lemma 8 (Optimal Cut Point τ). *Under the standard mixup scheme with $\mathbb{E}[W^2] = \mathbb{E}[W]$, the optimal cut point τ is 0.5. That is*

$$0.5 = \arg\min_{\tau \in \mathbb{R}} \left|u(W, \tau)\right|. \quad (14)$$

Lemma 8 shows that 0.5 minimizes $\left|u(W, \tau)\right|$. It is natural to employ $\tau = 0.5$ for the general-weight mixup as well because this cut point choice is more likely to preserve the conditional mean and conditional variance as shown in Corollary 5 and Theorem 6; we use $\tau = 0.5$ in what follows unless specified otherwise.

It is noted that categorical variable L in this paper can be regarded as a multinomial variable. Therefore, the above results concerning L are equally applicable for multiple categorical variables, since combinations of multiple multinomial variables also follow the multinomial distribution only with more individual categories.

4 STRUCTURE-PRESERVING MIXUP DISTRIBUTION

In this section we propose a new mixup scheme in data synthesis that can preserve key statistics such as the mean, variance and their conditional counterparts as discussed in Sec. 3.

4.1 Variance-Reduction Mixup

In the synthetic data literature the most common choice for the mixup weight is to use a distribution defined on $[0, 1]$ (Zhang et al., 2018; Verma et al., 2019; Cao et al., 2024). Two prominent examples are the $\mathrm{Beta}(\alpha, \beta)$ distribution with $\alpha, \beta \in (0, \infty)$ and the $\mathrm{Unif}(0, 1)$ distribution. Restricting the distribution's support to $[0, 1]$ does not distort the mean of the synthetic data, but it reduces the variance inevitably as shown below, which can be found in, e.g., Proposition 1 in Kim and Kim (2024).

Corollary 9 (Variance-Reduction Mixup). *For any synthetic variable $\tilde{X}$ generated by the mixup from a continuous X in (1), let the support of mixup weight variable W^X be bounded in $[0, 1]$. Then*

$$\mathrm{Var}[\tilde{X}] \leq \mathrm{Var}[X], \quad (15)$$

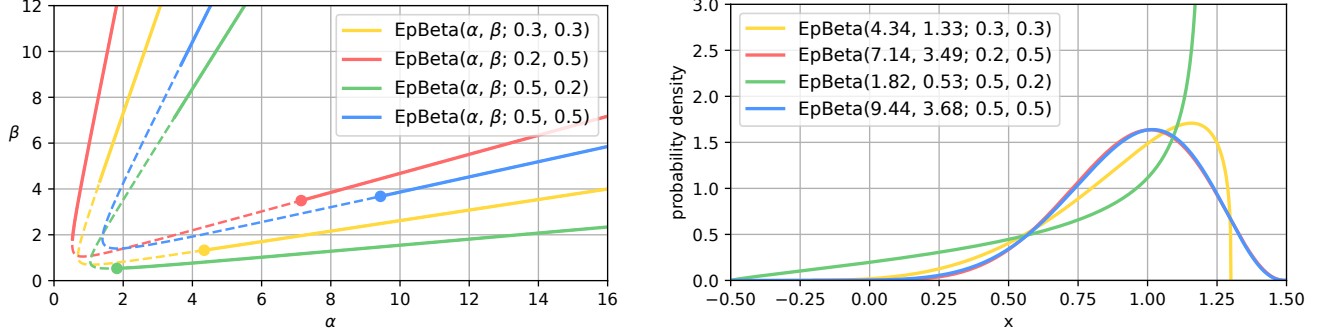*where the equality holds when $\Pr\{W^X \in \{0, 1\}\} = 1$.*

Figure 2: (Left) All (α, β) pairs that gives structure-preserving synthetic data for choices of (ϵ_0, ϵ_1). (Right) Density functions corresponding to the solid circled point of each curve on the left plot.

The condition $\Pr\{W^X \in \{0,1\}\} = 1$ implies that every synthetic instance is exactly same as one of the original instances, which essentially is a resampling scheme rather than data synthesis. Thus it is clear that the support of weight distribution must be expanded from $[0,1]$ in order to obtain variance-preserving synthetic data.

4.2 Structure-Preserving Mixup

The following simple example illustrates that the mixup weights generated from a constrained normal distribution with support $(-\infty, +\infty)$ can produce variance-preserving synthetic data.

Example 1. *Let the mixup weights are generated from the Gaussian distribution* $\mathrm{N}(\mu, \sigma^2)$ *where* $\sigma = \sqrt{\mu - \mu^2}$ *for some* $\mu \in [0,1]$, *i.e.,* $W^X = W^Y \sim \mathrm{N}(\mu, \mu - \mu^2)$. *Then, under the standard mixup scheme, we have, for any pair* (X,Y), $\mathrm{Var}[\tilde{X}] = \mathrm{Var}[X]$, $\mathrm{Var}[\tilde{Y}] = \mathrm{Var}[Y]$, *and* $\mathrm{Cov}[\tilde{X}, \tilde{Y}] = \mathrm{Cov}[X,Y]$.

A problem with this example however is that it can generate unacceptably extreme synthetic instances when the mixup weights take extreme values. One way to suppress extreme synthetic instances is to restrict the support of mixup weight distribution to a finite interval $[-\epsilon_0, 1 + \epsilon_1]$ for $\epsilon_0, \epsilon_1 \in [0, \infty)$, so that it always include $[0,1]$ as a sub-interval. To this extent we propose an expanded version of the standard beta distribution, also known as the four-parameter beta distribution (Johnson et al., 1995).

Definition 2 (Expanded Beta Distribution). *Let* $\epsilon_0, \epsilon_1 \in [0, \infty)$ *be given constants and* $V \sim \mathrm{Beta}(\alpha, \beta)$. *Then the random variable* W, *with support* $[-\epsilon_0, 1 + \epsilon_1]$, *is said to follow the expanded Beta distribution with parameters* $(\alpha, \beta, \epsilon_0, \epsilon_1)$, *or simply* $W \sim \mathrm{EpBeta}(\alpha, \beta; \epsilon_0, \epsilon_1)$, *if*

$$W = (1 + \epsilon_0 + \epsilon_1)V - \epsilon_0. \qquad (16)$$

The choice of (ϵ_0, ϵ_1) in practice would reflect the

meta-information or characteristics of the data. For example, if variable X cannot be negative, we should set the parameter as $\epsilon_0 = 0$ and $\epsilon_1 > 0$, with a further constraint $X_i \leq X_j$ for selected instances in (1) so that $\tilde{X}$ also remains positive. We also note that the size of ϵ_1 can control the maximum possible value of synthetic instances. To illustrate, consider an extreme case where the mixup weight is $1 + \epsilon_1$ coinciding with the upper bound, and the two selected original instances are $x_{\max} = \max_{i \in [n]}\{x_i\}$ and $x_{\min} = \min_{i \in [n]}\{x_i\}$. Then the resulting synthetic instance is $\tilde{x}_k = x_{\max} + \epsilon_1(x_{\max} - x_{\min})$, an extrapolation which leads to a much larger value than $x_{\max}$ by choosing a big ϵ_1.

After choosing (ϵ_0, ϵ_1) we can find some α, β that satisfy (17) in Theorem 10 and (18) in Theorem 11 to preserve data structure as follows.

Theorem 10. *For given* $\epsilon_0, \epsilon_1 \in [0, \infty)$, *consider an arbitrary synthetic pair* $(\tilde{X}, \tilde{Y})$ *generated from* (X, Y) *using the standard mixup scheme with* $W \sim \mathrm{EpBeta}(\alpha, \beta; \epsilon_0, \epsilon_1)$, *such that* $\alpha, \beta \in (0, \infty)$ *satisfy*

$$(1 + \epsilon_1 - \epsilon_0(\beta/\alpha)) \cdot (1 + \epsilon_0 - \epsilon_1(\alpha/\beta)) \cdot (1 + \alpha + \beta)$$
$$= (1 + \epsilon_0 + \epsilon_1)^2. \qquad (17)$$

Then we have $\mathrm{Var}[\tilde{X}] = \mathrm{Var}[X]$ *and* $\mathrm{Cov}[\tilde{X}, \tilde{Y}] = \mathrm{Cov}[X, Y]$.

Theorem 11. *Consider an arbitrary synthetic triple* $(\tilde{X}, \tilde{Y}, \tilde{L})$ *generated from* (X, Y, L) *using the standard mixup scheme with* $W \sim \mathrm{EpBeta}(\alpha, \beta; \epsilon_0, \epsilon_1)$ *for given* $\epsilon_0, \epsilon_1 \in [0, \infty)$ *and* $\tau = 0.5$. *Now suppose that, for a given* $\delta \in [0,1]$, (α, β) *satisfies (17) and the following*

$$\frac{1 + \epsilon_0 - \epsilon_1 \alpha/\beta}{1 + \alpha/\beta} + \frac{2(1 + \epsilon_0 + \epsilon_1)}{1 + \beta/\alpha} \frac{B(\tilde{\epsilon}; \alpha + 1, \beta)}{B(1; \alpha + 1, \beta)}$$
$$- (1 + 2\epsilon_0) \frac{B(\tilde{\epsilon}; \alpha, \beta)}{B(1; \alpha, \beta)} \leq \delta, \qquad (18)$$

where $B(x; \alpha, \beta) = \int_0^b t^{\alpha-1}(1-t)^{\beta-1} \, dt$ is the incomplete beta function and $\tilde{\epsilon} = \frac{0.5+\epsilon_0}{1+\epsilon_0+\epsilon_1}$.

Then, the gap of conditional (on categorical L) mean and variance are bounded as follows:

$$\left| \mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big] - \mathbb{E}\big[X\big|L=l\big] \right|$$
$$= \delta \cdot \Pr\{L \neq l\} \cdot \left| \mathbb{E}\big[X\big|L=l\big] - \mathbb{E}\big[X\big|L \neq l\big] \right|$$

and

$$\left| \mathrm{Var}\big[\tilde{X}\big|\tilde{L}=l\big] - \mathrm{Var}\big[X\big|L=l\big] \right|$$
$$\leq \delta \cdot \left| \mathrm{Var}\big[X\big|L=l\big] - \mathrm{Var}\big[X\big] \right|$$
$$+ \delta(1-\delta) \cdot \left(\mathbb{E}\big[X\big|L=l\big] - \mathbb{E}\big[X\big] \right)^2.$$

Theorem 11 implies that the magnitude of the gap can be controlled by δ. In particular, the conditional mean and variance can be perfectly preserved as δ tends to 0. This result, coupled with Theorem 10, suggests that δ can be viewed as a modulator that controls the amount of additional information to be borrowed from the original data.

On the left plot in Fig. 2, each curve represents (α, β) pairs that satisfy Theorem 10 for a selected (ϵ_0, ϵ_1) choice; each point on the curve therefore produces structure-preserving synthetic data. The solid-line part of each curve further satisfies Theorem 11 with equality for $\delta = 0.05$, so that the conditional mean and variance are almost preserved; for visual simplicity, we choose (α, β) pairs that make both sides of (18) equal with $\alpha \geq \beta$. The right plot in Fig. 2 shows the densities corresponding to each circled point shown on the left plot. For illustration Appendix E presents tables of EpBeta parameters (α, β) for additional values of ϵ_0, ϵ_1 and δ, all of which result in structure-preserving synthetic data.

We note that structure-preserving mixup weight distributions can be defined from other distributions in a similar manner, for example, the truncated normal distribution. However, these alternative weight distributions do not enjoy all the theoretical benefits that the EpBeta does, as shown in this section.

Guideline for Selecting Parameters. In practice, the user first specifies the possible ranges for ϵ_0 and ϵ_1, which determine the lower and upper bounds for the underlying distribution. Let $[x_l, x_u]$ be the conjectured bounds of the underlying distribution, satisfying $x_l \leq x_{\min}$ and $x_u \geq x_{\max}$. Under this assumption, both ϵ_0 and ϵ_1 are set to $\epsilon_0 = \epsilon_1 = \frac{x_u - x_l}{x_{\max} - x_{\min}} - 1$. Next, the user specifies δ to control the tolerance on differences in the conditional mean and variance. With these parameters fixed, the values of α and β are determined via Algorithm 1, which provides the weights

---

Algorithm 1 Mixup Weight from EpBeta

Input: $\epsilon_0, \epsilon_1 \geq 0$ (Smaller values better preserve the support of synthetic instances), $\delta \geq 0$ (Smaller values better preserve conditional mean and variance)
Output: w (Mixup weight)
1: Identify pairs (α, β) that satisfy the following constraints: $\alpha \geq \beta$, (17) in Theorem 10, and (18) in Theorem 11 for the given ϵ_0, ϵ_1, and δ.
2: Select the pair (α, β) for which α attains its minimum value.
3: Sample a Mixup weight w from the EpBeta$(\alpha, \beta; \epsilon_0, \epsilon_1)$ distribution.

---

used in mixup. An implementation of Algorithm 1 is available at: https://github.com/leechungpa/structure-preserving-mixup.

5 EXPERIMENTS

As in Sec 4, we propose the EpBeta distribution as a mixup weight distribution that more effectively preserves the original data distribution when generating synthetic data. In this section, we demonstrate the importance of this distribution not only for ensuring consistent statistical inference but also for maintaining model performance. First, we emphasize its significance for tabular data, which is highly structured. Then, we apply the proposed mixup to image datasets, showing that the structure-preserving synthetic data sustain the model performance under repeated data synthesis.

5.1 Tabular Data

We synthesize 6 different tabular datasets using three mixup methods (EpBeta, Beta$(0.1, 0.1)$, and Unif$(0, 1)$ and other four baseline methods available in open-source code (Qian et al., 2023); TVAE, CTGAN (Xu et al., 2019), TabDDPM (Kotelnikov et al., 2023), and GReaT (Borisov et al., 2023). The four number of EpBeta distribution parameter pairs have been selected so that it satisfies (17) and the equality condition (18) for given $\epsilon_0 = \epsilon_1 = 0.3$ and $\delta = 0.001, 0.005, 0.01$ or 0.05. These 10 synthetic datasets are evaluated and compared in terms of relative bias of key statistics, statistical inference, and the machine learning efficiency. Data descriptions and experimental details are in Appendix B.1.

Relative Bias. We compare the relative bias of covariance and expectation from each synthetic data, calculated as $\frac{\mathrm{Cov}[\tilde{X},\tilde{Y}] - \mathrm{Cov}[X,Y]}{\mathrm{Cov}[X,Y]}$ and $\frac{\mathbb{E}[\tilde{X}] - \mathbb{E}[X]}{\mathbb{E}[X]}$, respectively. As seen from Fig. 3, the covariance gets reduced when we use the Beta or Unif as a mixup
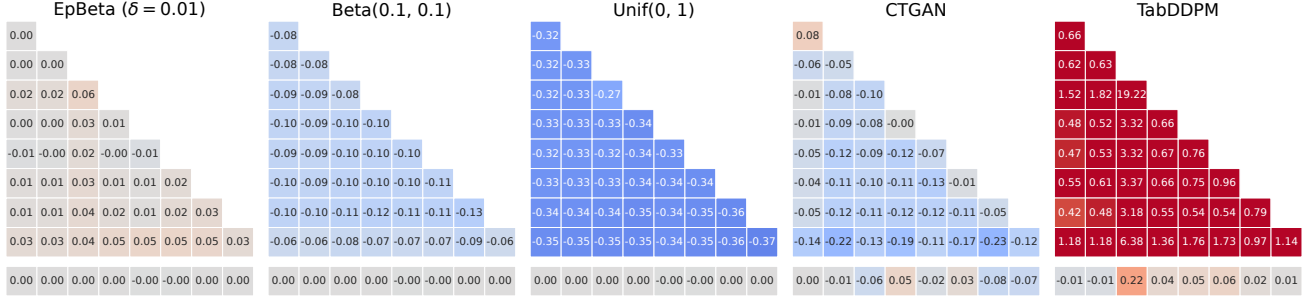
EpBeta ($\delta = 0.01$)

0.00							
0.00	0.00						
0.02	0.02	0.06					
0.00	0.00	0.03	0.01				
-0.01	-0.00	0.02	-0.00	-0.01			
0.01	0.01	0.03	0.01	0.01	0.02		
0.01	0.01	0.04	0.02	0.01	0.02	0.03	
0.03	0.03	0.04	0.05	0.05	0.05	0.05	0.03
0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00

Beta(0.1, 0.1)

-0.08							
-0.08	-0.08						
-0.09	-0.09	-0.08					
-0.10	-0.09	-0.10	-0.10				
-0.09	-0.09	-0.10	-0.10	-0.10			
-0.10	-0.09	-0.10	-0.10	-0.10	-0.11		
-0.10	-0.10	-0.11	-0.12	-0.11	-0.11	-0.13	
-0.06	-0.06	-0.08	-0.07	-0.07	-0.07	-0.09	-0.06
0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00

Unif(0, 1)

-0.32							
-0.32	-0.33						
-0.32	-0.33	-0.27					
-0.33	-0.33	-0.33	-0.34				
-0.32	-0.33	-0.32	-0.34	-0.33			
-0.33	-0.33	-0.33	-0.34	-0.34	-0.34		
-0.34	-0.34	-0.34	-0.35	-0.34	-0.35	-0.36	
-0.35	-0.35	-0.35	-0.35	-0.34	-0.35	-0.36	-0.37
0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00

CTGAN

0.08							
-0.06	-0.05						
-0.01	-0.08	-0.10					
-0.01	-0.09	-0.08	-0.00				
-0.05	-0.12	-0.09	-0.12	-0.07			
-0.04	-0.11	-0.10	-0.11	-0.13	-0.01		
-0.05	-0.12	-0.11	-0.12	-0.12	-0.11	-0.05	
-0.14	-0.22	-0.13	-0.19	-0.11	-0.17	-0.23	-0.12
0.00	-0.01	-0.06	0.05	-0.02	0.03	-0.08	-0.07

TabDDPM

0.66							
0.62	0.63						
1.52	1.82	19.22					
0.48	0.52	3.32	0.66				
0.47	0.53	3.32	0.67	0.76			
0.55	0.61	3.37	0.66	0.75	0.96		
0.42	0.48	3.18	0.55	0.54	0.54	0.79	
1.18	1.18	6.38	1.36	1.76	1.73	0.97	1.14
-0.01	-0.01	0.22	0.04	0.05	0.06	0.02	0.01

Figure 3: The relative bias of covariance (triangle) and expectation (bar at bottom) for the 'Abalone' data using various synthetic generation methods. Negative bias is colored in blue and positive bias in red. Grey represents bias close to zero.

weight distribution, confirming Corollary 9. In contrast, the ML-based synthetic datasets (CTGAN and TabDDPM) substantially disturb the covariance and expectation. The proposed EpBeta generates the most balanced synthetic data; the results of the other data are presented in Appendix B.2. This exercise shows that common data synthesis techniques are subject to considerable distortion in basic distributional quantities, which are often important in the early data exploration stage.
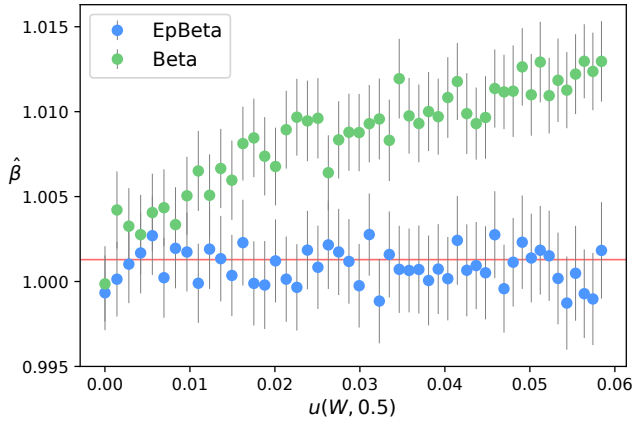

Figure 4: The estimated polynomial regression coefficient and its 95% confidence interval for each synthetic data. The red horizontal line represents the coefficient estimate from the original data.

Statistical Inference. We present a simple polynomial regression example that illustrates why preserving the variance is important from a statistical perspective. The regression analysis is a widely used tool in various studies. Specifically researchers are interested in accurate estimation of regression coefficients, which relate the target variable to the other attributes and allow an appropriate interpretation about their relationship. As previously mentioned, synthetic data generated by any mixup weight distribution preserves the coefficient of linear regression, as long as no poly-

nomial terms are involved. For our experiment we draw 1000 original instances with two variables (X, Y) where $X \sim \mathrm{N}(5, 1)$ and $Y \sim \mathrm{N}(X^2, 1)$. Then we fit a quadratic regression $Y = \beta X^2 + e$ where $e \sim \mathrm{N}(0, \sigma^2)$, and obtain the coefficient estimate $\hat{\beta}$. From this original data, we generate synthetic data $(\tilde{X}, \tilde{Y})$ with same number of instances using, respectively, EpBeta and Beta as the weight distribution, which have various $u(W, 0.5)$ values. For each synthetic data we repeat the same regression fitting; if the synthesis is successful, the resulting coefficient estimates should be close to $\hat{\beta}$, the coefficient obtained from the original data. The result is presented in Fig. 4 which shows that the proposed EpBeta consistently produces coefficient estimates close to the true value $\hat{\beta}$, whereas Beta weight distribution yields coefficient estimates that are sometimes unacceptably distant from the true value, with $\hat{\beta}$ sitting outside the confidence interval. This example elucidates that a data synthesis method that does not preserve the variance can fail even in rudimentary statistical analyses. We mention that EpBeta may produce biased estimators in different setting, but the bias can be reduced by using a small enough $\delta \in [0, 1]$.

In Appendix C, we also include a classification example to highlight the significance of statistical inference.

Machine Learning Efficiency. To compare the machine learning efficiency we use synthetic datasets to train various models, such as CatBoost (Prokhorenkova et al., 2018) and MLP, following the experiment protocol of Gorishniy et al. (2021); Zhao et al. (2021); Kotelnikov et al. (2023). In particular, each model has been trained using one of the synthetic datasets and tested against the original data. The focus here is on assessing how closely each synthetic data resembles the original data, rather than on effectiveness of the models. Tables in Appendix B.3 show that the performance of the mixup-driven synthetic datasets is comparable to other ML-based synthetic methods.

5.2 Image Data

We synthesize image data using the mixup method, and demonstrate that preserving data structure can prevent model collapse, emphasizing its significance.

Model Collapse with Repeated Synthesis. Recent work in Nature discusses *model collapse*, a phenomenon where models trained predominantly on synthetic data begin to forget rare information, leading to significant performance degradation as human-generated data becomes scarce (Shumailov et al., 2024). They showed that training a language model with texts, followed by training a new model on the synthetic texts generated from the previously learned model over nine iterations, results in reduced performance.

We conduct similar experiments in the image domain to show that using EpBeta distribution is more effective at preventing model collapse compared to the original mixup which reduces variance. We use the mixup method on the CIFAR-10 dataset (Krizhevsky et al., 2009) to create synthetic images and repeatedly synthesized from these generated images, using the EpBeta distribution with $\epsilon_0 = \epsilon_1 = 0.3$ and $\delta = 0.05$, and the $\mathrm{Unif}(0,1)$ distribution, respectively. We then train Resnet-18 on the synthesized images to classify the image labels, and evaluate top-1 accuracy on the original test set.

The baseline model, trained on the original dataset, achieves a top-1 accuracy of 79.55. As demonstrated in Table 1, training with synthetic datasets generated using either the EpBeta or Unif distributions enhances model performance when the resynthesis process is limited to 10 iterations or fewer. However, beyond 20 iterations of resynthesis, the EpBeta distribution maintains consistent performance by preserving the original data structure, while the $\mathrm{Unif}(0,1)$ distribution results in substantial performance degradation. These findings suggest that structure-preserving synthetic data generation enables sustained model performance and mitigates the risk of model collapse.

Table 1: Top-1 accuracy of image classification models trained on repeatedly synthesized data. Each cell reports the mean and standard deviation of top-1 accuracy across five independently trained models, each using distinct randomly generated synthetic datasets.

Resynthesis	5	10	15	20	25
EpBeta($\delta = 0.05$)	85.78 (0.13)	85.99 (0.21)	86.08 (0.28)	86.43 (0.14)	85.76 (0.26)
Unif$(0,1)$	84.39 (0.29)	83.83 (0.13)	74.49 (0.53)	21.81 (4.21)	12.34 (1.07)

6 CONCLUSION

This paper presents significant theoretical advancements in the context of the mixup method for data synthesis. With a focus on ensuring that synthetic data mirrors the original in all aspects, the primary contribution is the establishment of specific conditions that the mixup weight distribution must meet to preserve the original data's structure, including its mean, variance, and their conditional counterparts. On the theoretical front, we derive conditions under which the (co)variance for any pair of continuous variables remains intact with mixup. Additionally, we prove that the mean and variance, when conditioned on a categorical variable, can be preserved within a defined error bound. To achieve this, we introduce a class of mixup weight distributions, called EpBeta, a generalized form of the Beta distribution, which adheres to these theoretical conditions, thereby preserving the structural integrity of the original data. Our numerical experiments confirm that our proposed mixup method maintains essential distributional properties, leading to more accurate statistical inferences. In terms of performance, ours delivers results comparable to other synthetic data generation methods while significantly maintaining performance under repeated synthesis.

While our method primarily focuses on the classical mixup framework, its applicability extends to related data augmentation techniques such as CutMix (Yun et al., 2019), where mixup weights take binary values (0 or 1), and to its extensions like Gaussian-Mixup (Park et al., 2022b) which replaces discrete region constraints with continuous distributions. Exploring optimal parameterizations for such distributions remains a research direction that could further enhance the flexibility of the mixup methods.

In addition to these extensions, future research will explore non-equal weight mixup methods that aim to identify representative instances rather than selecting them uniformly at random. Moreover, we may consider more flexible weight distribution classes to further improve the quality of synthetic data while preserving essential statistical properties.

References

Baena, R., Drumetz, L., and Gripon, V. (2022). Preventing manifold intrusion with locality: Local mixup. *arXiv preprint arXiv:2201.04368*.

Baowaly, M. K., Lin, C.-C., Liu, C.-L., and Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. (2023). Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., and Rueckert, D. (2018). Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2012). Dbsmote: density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36:664–684.

Cao, C., Zhou, F., Dai, Y., Wang, J., and Zhang, K. (2024). A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *ACM Computing Surveys*, 57(2):1–38.

Carratino, L., Cissé, M., Jenatton, R., and Vert, J.-P. (2022). On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, J., Wang, Z., Tian, R., Yang, Z., and Yang, D. (2020). Local additivity based data augmentation for semi-supervised NER. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1241–1251, Online. Association for Computational Linguistics.

Chen, M., Fu, D. Y., Narayan, A., Zhang, M., Song, Z., Fatahalian, K., and Ré, C. (2022). Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pages 3090–3122. PMLR.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613.

Douzas, G., Bacao, F., and Last, F. (2018). Improving imbalanced learning through a heuristic over-sampling method based on k-means and smote. *Information sciences*, 465:1–20.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943.

Gulati, M. and Roysdon, P. (2024). Tabmt: Generating tabular data with masked transformers. *Advances in Neural Information Processing Systems*, 36.

Guo, H. (2020). Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4044–4051.

Guo, H., Mao, Y., and Zhang, R. (2019). Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3714–3722.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Radke, R., and Feris, R. (2021). A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8845–8855.

Johnson, B. A., Tateishi, R., and Hoan, N. T. (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International journal of remote sensing*, 34(20):6969–6982.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Kim, B. J. and Kim, S. W. (2024). Configuring data augmentations to reduce variance shift in positional embedding of vision transformers. *arXiv preprint arXiv:2405.14115*.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386.

Kim, J., Lee, C., and Park, N. (2023). STaSy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*.

Kim, J., Lee, C., Shin, Y., Park, S., Kim, M., Park, N., and Cho, J. (2022). Sos: Score-based oversampling for tabular data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 762–772.

Kim, Y.-Y., Song, K., Jang, J., and Moon, I.-C. (2021). Lada: Look-ahead data acquisition via augmentation for deep active learning. *Advances in Neural Information Processing Systems*, 34:22919–22930.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lee, C., Kim, J., and Park, N. (2023). Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pages 18940–18956. PMLR.

Lee, C., Oh, J., Lee, K., and Sohn, J.-y. (2025). A theoretical framework for preventing class collapse in supervised contrastive learning. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*.

Lee, H., Lee, K., Lee, K., Lee, H., and Shin, J. (2021). Improving transferability of representations via augmentation-aware self-supervision. *Advances in Neural Information Processing Systems*, 34:17710–17722.

Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M., and Zhang, C. (2020). Vaem: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33:11237–11247.

Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., and Ford, W. (1995). Abalone. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55C7W.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729. IEEE.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26.

Park, C., Yun, S., and Chun, S. (2022a). A unified analysis of mixed sample data augmentation: A loss function perspective. *Advances in Neural Information Processing Systems*, 35:35504–35518.

Park, C., Yun, S., and Chun, S. (2022b). A unified analysis of mixed sample data augmentation: A loss function perspective. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35504–35518. Curran Associates, Inc.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.*, 11(10):1071–1083.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Qian, Z., Cebere, B.-C., and van der Schaar, M. (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv:2301.07573*.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Si, Y. and Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of educational and behavioral statistics*, 38(5):499–521.

Sohn, J.-Y., Shang, L., Chen, H., Moon, J., Papailiopoulos, D., and Lee, K. (2022). GenLabel: Mixup relabeling using generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20278–20313. PMLR.

Solatorio, A. V. and Dupriez, O. (2023). Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*.

Sun, K., Yu, B., Lin, Z., and Zhu, Z. (2024). Patch-level neighborhood interpolation: A general and effective graph-based regularization strategy. In *Asian Conference on Machine Learning*, pages 1276–1291. PMLR.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhang, H., Zhang, J., Shen, Z., Srinivasan, B., Qin, X., Faloutsos, C., Rangwala, H., and Karypis, G. (2024). Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. (2021). How does mixup help with robustness and generalization? In *International Conference on Learning Representations*.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. (2022a). When and how mixup improves calibration. In *International Conference on Machine Learning*, pages 26135–26160. PMLR.

Zhang, S., Liu, M., Yan, J., Zhang, H., Huang, L., Yang, X., and Lu, P. (2022b). M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2461–2470.

Zhang, T., Wang, S., Yan, S., Jian, L., and Liu, Q. (2023). Generative table pre-training empowers models for tabular prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14836–14854, Singapore. Association for Computational Linguistics.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. Y. (2021). Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR.

Zhao, Z., Kunar, A., Birke, R., Van der Scheer, H., and Chen, L. Y. (2024). Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data*, 6:1296508.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sec. 3.1.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Sec. 3.1.

   (b) Complete proofs of all theoretical results. [Yes] See Appendix A.

   (c) Clear explanations of any assumptions. [Yes] See Sec. 3.1.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Sec. 5 and Appendix B. An implementation of Algorithm 1 is available at: https://github.com/leechungpa/structure-preserving-mixup.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Sec. 5 and Appendix B.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Sec. 5 and Appendix B.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Appendix B.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] See Sec. 5 and Appendix B.

   (b) The license information of the assets, if applicable. [Yes] See Table 2.

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS

As expectation is preserved, the distribution of synthetic categorical variable is also preserved, by following a simple theorem.

Theorem A.1 (Category-Preserving Mixup). *For any synthetic variable $\tilde{L}$ generated from L in* (3), *the distribution is preserved. That is* $\Pr\{L = l\} = \Pr\{\tilde{L} = l\}$ *for any* l.

Proof. It can be directly proved by the definition in (3), as follow.

$$
\begin{aligned}
\Pr\{\tilde{L} = z\} &= \Pr\{L_i = l\}\Pr\{W^L \geq \tau\} + \Pr\{L_j = l\}\Pr\{W^L < \tau\} \\
&= \Pr\{L = l\}\big(\Pr\{W^L \geq \tau\} + \Pr\{W^L < \tau\}\big) \\
&= \Pr\{L = l\}
\end{aligned}
$$

$\square$

A.1 Regression Coefficients Are Preserved

Theorem A.2. *For any synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the standard mixup scheme, the regression coefficients are preserved. I.e.,*

$$
\arg\min_{\beta_0, \beta_1} \mathbb{E}\big\|Y - \beta_0 - \beta_1 X\big\|_2^2 = \arg\min_{\beta_0, \beta_1} \mathbb{E}\big\|\tilde{Y} - \beta_0 - \beta_1 \tilde{X}\big\|_2^2.
$$

Proof. It is well known that $\hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \mathbb{E}[X]$ and $\hat{\beta}_1 = \mathrm{Corr}[X, Y]$ from the convex optimization:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \beta_0} \mathbb{E}\big\|Y - \beta_0 - \beta_1 X\big\|_2^2 \bigg|_{(\beta_0, \beta_1) = (\hat{\beta}_0, \hat{\beta}_1)} \\
&= 2\hat{\beta}_0 - 2\mathbb{E}[Y] - \hat{\beta}_1 \mathbb{E}[X], \\
0 &= \frac{\partial}{\partial \beta_1} \mathbb{E}\big\|Y - \beta_0 - \beta_1 X\big\|_2^2 \bigg|_{(\beta_0, \beta_1) = (\hat{\beta}_0, \hat{\beta}_1)} \\
&= 2\mathbb{E}[X^2]\hat{\beta}_1 - 2\mathbb{E}[X(Y - \hat{\beta}_0)] \\
&= 2\mathbb{E}[(X - \mathbb{E}[X])^2]\hat{\beta}_1 - 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].
\end{aligned}
$$

By Corollary A.5, the correlation is preserved under the standard mixup scheme. Moreover, expectation is always preserved. Therefore,

$$
\begin{aligned}
\arg\min_{\beta_0, \beta_1} \mathbb{E}\big\|Y - \beta_0 - \beta_1 X\big\|_2^2 &= \big(\mathbb{E}[Y] - \mathrm{Corr}[X, Y]\mathbb{E}[X],\ \mathrm{Corr}[X, Y]\big) \\
&= \big(\mathbb{E}[\tilde{X}] - \mathrm{Corr}[\tilde{X}, \tilde{Y}]\mathbb{E}[\tilde{X}],\ \mathrm{Corr}[\tilde{X}, \tilde{Y}]\big) \\
&= \arg\min_{\beta_0, \beta_1} \mathbb{E}\big\|\tilde{Y} - \beta_0 - \beta_1 \tilde{X}\big\|_2^2.
\end{aligned}
$$

This can be easily generalized to multiple linear regression. For simplicity, assume that the variables are centered. Define $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$, and $\beta \in \mathbb{R}^{p \times 1}$ where $\hat{\beta} = \arg\min_\beta \mathbb{E}\big\|\mathbf{Y} - \mathbf{X}\beta\big\|_2^2$. It is well known that $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$. Then, the coefficients are preserved as below,

$$
\begin{aligned}
\hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} \\
&= \big((\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\big)^{-1} \cdot (\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]) \\
&= \Big(\big(1 + 2\mathbb{E}[(W^X)^2 - W^X]\big)(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\Big)^{-1} \cdot \big(1 + 2\mathbb{E}[(W^X)^2 - W^X]\big)(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])
\end{aligned}
$$

$$
\tag{19}
$$

$$
= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}
$$

where the equality in (19) holds from Corollary A.5. These results can be easily extended to the non-centered case when an intercept term is included in the regression, because mixup always preserves the expectation. $\square$

A.2 Proofs for Sec. 3

Lemma A.3 (Variance)**.** *For any synthetic $\tilde{X}$ generated from X in (1):*

1. $\mathrm{Var}\big[\tilde{X}\big] = \mathrm{Var}\big[X\big]$, *if and only if the first and second moments of mixup weight are equal. That is*

$$\mathbb{E}\big[(W^X)^2\big] = \mathbb{E}\big[W^X\big]. \tag{20}$$

2. $\mathrm{Var}\big[\tilde{X}\big] < \mathrm{Var}\big[X\big]$, *if and only if $\mathbb{E}\big[(W^X)^2\big] < \mathbb{E}\big[W^X\big]$.*

3. $\mathrm{Var}\big[\tilde{X}\big] > \mathrm{Var}\big[X\big]$. *if and only if $\mathbb{E}\big[(W^X)^2\big] > \mathbb{E}\big[W^X\big]$.*

Proof. Since $\mathbb{E}\big[\tilde{X}\big] = \mathbb{E}\big[X\big]$, the variance of synthetic data $\mathrm{Var}\big[\tilde{X}\big]$ is determined as

$$
\begin{aligned}
\mathrm{Var}\big[\tilde{X}\big] &= \mathbb{E}\big[W^X X_i + (1 - W^X) X_j\big]^2 - \mathbb{E}\big[X\big]^2 \\
&= \mathbb{E}\big[(W^X)^2\big]\mathbb{E}\big[X_i^2\big] + \mathbb{E}\big[(1 - W^X)^2\big]\mathbb{E}\big[X_j^2\big] + 2\mathbb{E}\big[W^X(1 - W^X)\big]\mathbb{E}\big[X_i\big]\mathbb{E}\big[X_j\big] - \mathbb{E}\big[X\big]^2 \\
&= \mathbb{E}\big[(W^X)^2 + (1 - W^X)^2\big]\big(\mathrm{Var}\big[X\big] + \mathbb{E}\big[X\big]^2\big) + 2\mathbb{E}\big[W^X(1 - W^X)\big]\mathbb{E}\big[X\big]^2 - \mathbb{E}\big[X\big]^2 \\
&= \mathrm{Var}\big[X\big] + 2\mathbb{E}\big[(W^X)^2 - W^X\big]\mathrm{Var}\big[X\big]. \tag{21}
\end{aligned}
$$

Therefore, $\mathrm{Var}\big[\tilde{X}\big] = \mathrm{Var}\big[X\big]$ holds if and only if $\mathbb{E}\big[(W^X)^2 - W^X\big] = 0$. The cases of variance reduction and inflation can be trivially shown using a similar approach. □

Theorem A.4 (Covariance)**.** *For any synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the general-weight mixup in (1) and (2):*

1. $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Cov}\big[X, Y\big]$, *if and only if*

$$\mathbb{E}\big[W^X W^Y\big] = \frac{1}{2}\big(\mathbb{E}\big[W^X\big] + \mathbb{E}\big[W^Y\big]\big). \tag{22}$$

2. $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] < \mathrm{Cov}\big[X, Y\big]$, *if and only if $\mathbb{E}\big[W^X W^Y\big] < \frac{1}{2}\big(\mathbb{E}\big[W^X\big] + \mathbb{E}\big[W^Y\big]\big)$.*

3. $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] > \mathrm{Cov}\big[X, Y\big]$, *if and only if $\mathbb{E}\big[W^X W^Y\big] > \frac{1}{2}\big(\mathbb{E}\big[W^X\big] + \mathbb{E}\big[W^Y\big]\big)$.*

Proof. Since $\mathbb{E}\big[\tilde{X}\big] = \mathbb{E}\big[X\big]$ and $\mathbb{E}\big[\tilde{Y}\big] = \mathbb{E}\big[Y\big]$, the covariance of synthetic data $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big]$ is determined as

$$
\begin{aligned}
\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] &= \mathbb{E}\big[\tilde{X}\tilde{Y}\big] - \mathbb{E}\big[\tilde{X}\big]\mathbb{E}\big[\tilde{Y}\big] \\
&= \mathbb{E}\big[W^X W^Y X_i Y_i + (1 - W^X) W^Y X_j Y_i\big]\mathbb{E}\big[W^X(1 - W^Y)X_i Y_j + (1 - W^X)(1 - W^Y)X_j Y_j\big] \\
&\quad - \mathbb{E}\big[X\big]\mathbb{E}\big[Y\big] \\
&= \mathbb{E}\big[2W^X W^Y + 1 - W^X - W^Y\big]\mathbb{E}\big[X_i Y_i\big] - \mathbb{E}\big[2W^X W^Y + 1 - W^X - W^Y\big]\mathbb{E}\big[X\big]\mathbb{E}\big[Y\big] \\
&= \mathbb{E}\big[2W^X W^Y + 1 - W^X - W^Y\big]\mathrm{Cov}\big[X, Y\big] \\
&= \mathrm{Cov}\big[X, Y\big] + \mathbb{E}\big[2W^X W^Y - W^X - W^Y\big]\mathrm{Cov}\big[X, Y\big]. \tag{23}
\end{aligned}
$$

Therefore, $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Cov}\big[X, Y\big]$ if and only if $\mathbb{E}\big[2W^X W^Y - W^X - W^Y\big] = 0$, which is equal to Eq. (6). The cases of covariance reduction and inflation can be trivially shown using a similar approach. □

Corollary A.5 (Correlation)**.** *For any synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the standard mixup scheme, we have*

$$\mathrm{Corr}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Corr}\big[X, Y\big].$$

Proof. From (23) in Theorem A.4 with using the standard mixup scheme, i.e., $W_k^X = W_k^Y$ for all $k \in [m]$, the covariance of synthetic data $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big]$ is determined as

$$\begin{aligned}
\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] &= \mathrm{Cov}\big[X, Y\big] + \mathbb{E}\big[2W^X W^Y - W^X - W^Y\big]\mathrm{Cov}\big[X, Y\big] \\
&= \mathrm{Cov}\big[X, Y\big] + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\mathrm{Cov}\big[X, Y\big] \\
&= \big(1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\big)\mathrm{Cov}\big[X, Y\big].
\end{aligned}$$

As a result, using (21) in Lemma A.3, the correlation of synthetic data is preserved as

$$\begin{aligned}
\mathrm{Corr}\big[\tilde{X}, \tilde{Y}\big] &= \frac{\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big]}{\sqrt{\mathrm{Var}\big[\tilde{X}\big]\mathrm{Var}\big[\tilde{Y}\big]}} \\
&= \frac{\big(1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\big)\mathrm{Cov}\big[X, Y\big]}{\sqrt{\big(1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\big)^2 \mathrm{Var}\big[X\big]\mathrm{Var}\big[Y\big]}} \\
&= \frac{\big(1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\big)\mathrm{Cov}\big[X, Y\big]}{\big|1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big]\big|\sqrt{\mathrm{Var}\big[X\big]\mathrm{Var}\big[Y\big]}} \\
&= \mathrm{Corr}\big[X, Y\big].
\end{aligned}$$

Note that the last equality comes from the fact that

$$1 + 2\mathbb{E}\big[\big(W^X\big)^2 - W^X\big] = \mathbb{E}\big[\big(W^X\big)^2\big] + \mathbb{E}\big[\big(W^X - 1\big)^2\big] \geq 0.$$

$\square$

Theorem A.6 (Conditional Mean)**.** *For any synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup, where X is continuous and L is categorical, the synthetic conditional mean $\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big]$ can be expressed as*

$$\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] = \big(1 - u\big(W^X, W^L, \tau\big)\big)\mathbb{E}\big[X|L = l\big] + u\big(W^X, W^L, \tau\big)\mathbb{E}\big[X\big],$$

or, alternatively,

$$\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] = \big(1 - u\big(W^X, W^L, \tau\big)\Pr\{L \neq l\}\big)\mathbb{E}\big[X|L = l\big] + u\big(W^X, W^L, \tau\big)\Pr\{L \neq l\}E[X|L \neq l].$$

Corollary A.7 (Conditional Mean Gap)**.** *For any synthetic pair $(\tilde{X}, \tilde{L})$ generated from (X, L) using the general-weight mixup, where X is continuous and L is categorical, the difference between the conditional mean is given by*

$$\big|\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathbb{E}\big[X|L = l\big]\big| = \big|u\big(W^X, W^L, \tau\big)\big| \cdot \Pr\{L \neq l\} \cdot \big|\mathbb{E}\big[X|L = l\big] - \mathbb{E}\big[X|L \neq l\big]\big|. \tag{24}$$

Proof of Theorem A.6 and Corollary A.7. Let us define the sets $A_1 = \{L_i = l, W^L \geq \tau\}$ and $A_2 = \{L_j = l, W^L < \tau\}$, which disjointly divide the set $\{\tilde{L} = l\} = A_1 \dot\cup A_2$. Then, $\Pr(A_1) = \Pr\{L_i = l\}\Pr\{W^L \geq \tau\}$ and $\Pr(A_2) = \Pr\{L_j = l\}\Pr\{W^L < \tau\}$.

To obtain $\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] = \mathbb{E}\big[\tilde{X}\big|A_1 \cup A_2\big]$, we first determine $\mathbb{E}\big[\tilde{X}\big|A_1\big]$ and $\mathbb{E}\big[\tilde{X}\big|A_2\big]$ as below:

$$\begin{aligned}
\mathbb{E}\big[\tilde{X}\big|A_1\big] &= \mathbb{E}\big[W^X X_i + (1 - W^X)X_j\big|\{L_i = l, W^L \geq \tau\}\big] \\
&= \mathbb{E}\big[W^X\big|W^L \geq \tau\big]\mathbb{E}\big[X_i\big|L_i = l\big] + \big(1 - \mathbb{E}\big[W^X\big|W^L \geq \tau\big]\big)\mathbb{E}\big[X_j\big] \\
&= \mathbb{E}\big[W^X\big|W^L \geq \tau\big]\mathbb{E}\big[X|L = l\big] + \big(1 - \mathbb{E}\big[W^X\big|W^L \geq \tau\big]\big)\mathbb{E}\big[X\big], \\
\mathbb{E}\big[\tilde{X}\big|A_2\big] &= \mathbb{E}\big[W^X\big|W^L < \tau\big]\mathbb{E}\big[X\big] + \big(1 - \mathbb{E}\big[W^X\big|W^L < \tau\big]\big)\mathbb{E}\big[X|L = l\big].
\end{aligned}$$

Therefore, the mean of the continuous variable conditioned by the categorical variable from the synthetic data is

$$
\begin{aligned}
\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big] &= \mathbb{E}\big[\tilde{X}\big|A_1 \cup A_2\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\tilde{X}\big|A_i\big]\big|A_1 \cup A_2\big] \\
&= \mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\big[A_1\big|A_1 \cup A_2\big] + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\big[A_2\big|A_1 \cup A_2\big] \\
&= \frac{\mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\big[A_1\big] + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\big[A_2\big]}{Pr\big[A_1 \cup A_2\big]} \\
&= \frac{\mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\big[A_1\big] + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\big[A_2\big]}{\Pr\big[A_1\big] + \Pr\big[A_2\big]} \\
&= \frac{\mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\{L_i=l\}\Pr\{W^L \geq \tau\} + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\{L_j=l\}\Pr\{W^L < \tau\}}{\Pr\{L_i=l\}\Pr\{W^L \geq \tau\} + \Pr\{L_j=l\}\Pr\{W^L < \tau\}} \\
&= \frac{\mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\{W^L \geq \tau\} + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\{W^L < \tau\}}{\Pr\{W^L \geq \tau\} + \Pr\{W^L < \tau\}} \\
&= \mathbb{E}\big[\tilde{X}\big|A_1\big]\Pr\{W^L \geq \tau\} + \mathbb{E}\big[\tilde{X}\big|A_2\big]\Pr\{W^L < \tau\} \\
&= \big(\mathbb{E}[W^X|W^L \geq \tau]\mathbb{E}[X|L=l] + \big(1 - \mathbb{E}[W^X|W^L \geq \tau]\big)\mathbb{E}[X]\big)\Pr\{W^L \geq \tau\} \\
&\quad + \big(\mathbb{E}[W^X|W^L < \tau]\mathbb{E}[X] + \big(1 - \mathbb{E}[W^X|W^L < \tau]\big)\mathbb{E}[X|L=l]\big)\big(1 - \Pr\{W^L \geq \tau\}\big) \\
&= \big(1 - u\big(W^X, W^L, \tau\big)\big)\mathbb{E}[X|L=l] + u\big(W^X, W^L, \tau\big)\mathbb{E}[X],
\end{aligned}
$$

where $u\big(W^X, W^L, \tau\big)$ is defined as

$$
\begin{aligned}
u\big(W^X, W^L, \tau\big) &= \Pr\{W^L \geq \tau\} - \mathbb{E}\big[W^X\big|W^L \geq \tau\big]\Pr\{W^L \geq \tau\} \\
&\quad + \mathbb{E}\big[W^X\big|W^L < \tau\big] - \mathbb{E}\big[W^X\big|W^L < \tau\big]\Pr\{W^L \geq \tau\} \\
&= \mathbb{E}\big[1 - W^X\big|W^L \geq \tau\big]\Pr\{W^L \geq \tau\} + \mathbb{E}\big[W^X\big|W^L < \tau\big]\Pr\{W^L < \tau\} \\
&= \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^L \geq \tau\}\big] + \mathbb{E}\big[W^X\mathbf{I}\{W^L < \tau\}\big] \\
&= \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\}\big],
\end{aligned}
$$

and $\mathbf{I}$ is an indicator function.

Moreover, the law of total expectation,

$$
E[X] = E[E[X|Y]] = E[X|L=l]\Pr\{L=l\} + E[X|L \neq l]\Pr\{L \neq l\},
$$

implies that

$$
\begin{aligned}
\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big] &= \big(1 - u\big(W^X, W^L, \tau\big)\big)\mathbb{E}[X|L=l] + u\big(W^X, W^L, \tau\big)E[X|L=l]\Pr\{L=l\} \\
&\quad + u\big(W^X, W^L, \tau\big)E[X|L \neq l]\Pr\{L \neq l\} \\
&= \big(1 - u\big(W^X, W^L, \tau\big)\Pr\{L \neq l\}\big)\mathbb{E}[X|L=l] + u\big(W^X, W^L, \tau\big)\Pr\{L \neq l\}E[X|L \neq l],
\end{aligned}
$$

which is equal to

$$
\big|\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big] - \mathbb{E}\big[X\big|L=l\big]\big| = \big|u\big(W^X, W^L, \tau\big)\big| \cdot \Pr\{L \neq l\} \cdot \big|\mathbb{E}\big[X\big|L=l\big] - \mathbb{E}\big[X\big|L \neq l\big]\big|.
$$

$\square$

Theorem A.8 (Conditional Variance Gap). *Assume that* $\mathbb{E}\big[(W^X)^2\big] = \mathbb{E}\big[W^X\big]$. *Then, for any synthetic pair* $(\tilde{X}, \tilde{L})$ *generated from* (X, L) *using the general-weight mixup, where* X *is continuous and* L *is categorical, the difference between the conditional variance is bounded as follows:*

$$\Big|\text{Var}\big[\tilde{X}\big|\tilde{L} = l\big] - \text{Var}\big[X\big|L = l\big]\Big| \leq \big|u(W^X, W^L, \tau)\big| \cdot \Big|\text{Var}\big[X\big|L = l\big] - \text{Var}\big[X\big]\Big|$$
$$+ \big|u(W^X, W^L, \tau)(1 - u(W^X, W^L, \tau))\big| \cdot \big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2.$$

Proof. Let us define the sets $A_1 = \{L_i = l, W^L \geq \tau\}$ and $A_2 = \{L_j = l, W^L < \tau\}$, which disjointly divide the set $\{\tilde{L} = l\} = A_1 \dot{\cup} A_2$. Then, $\Pr(A_1) = \Pr\{L_i = l\}\Pr\{W^L \geq \tau\}$ and $\Pr(A_2) = \Pr\{L_j = l\}\Pr\{W^L < \tau\}$.

To obtain $\mathbb{E}\big[\tilde{X}^2\big|\tilde{L} = l\big] = \mathbb{E}\big[\tilde{X}^2\big|A_1 \cup A_2\big]$, we first determine $\mathbb{E}\big[\tilde{X}^2\big|A_1\big]$ and $\mathbb{E}\big[\tilde{X}^2\big|A_2\big]$ as follows:

$$\mathbb{E}\big[\tilde{X}^2\big|A_1\big] = \mathbb{E}\big[(W^X)^2 X_i^2 + (1 - W^X)^2 X_j^2 + 2W^X(1 - W^X)X_i X_j\big|\{L_i = l, W^L \geq \tau\}\big]$$
$$= \mathbb{E}\big[(W^X)^2\big|W^L \geq \tau\big]\mathbb{E}\big[X_i^2\big|L_i = l\big] + \mathbb{E}\big[(1 - W^X)^2\big|W^L \geq \tau\big]\mathbb{E}\big[X_j^2\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\big|W^L \geq \tau\big]\mathbb{E}\big[X_i\big|L_i = l\big]\mathbb{E}\big[X_j\big]$$
$$= \mathbb{E}\big[(W^X)^2\big|W^L \geq \tau\big]\mathbb{E}\big[X^2\big|L = l\big] + \mathbb{E}\big[(1 - W^X)^2\big|W^L \geq \tau\big]\mathbb{E}\big[X^2\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\big|W^L \geq \tau\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big],$$
$$\mathbb{E}\big[\tilde{X}^2\big|A_2\big] = \mathbb{E}\big[(W^X)^2 X_i^2 + (1 - W^X)^2 X_j^2 + 2W^X(1 - W^X)X_i X_j\big|\{L_j = l, W^L < \tau\}\big]$$
$$= \mathbb{E}\big[(W^X)^2\big|W^L < \tau\big]\mathbb{E}\big[X^2\big] + \mathbb{E}\big[(1 - W^X)^2\big|W^L < \tau\big]\mathbb{E}\big[X^2\big|L = l\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\big|W^L < \tau\big]\mathbb{E}\big[X\big]\mathbb{E}\big[X\big|L = l\big].$$

Then,

$$\mathbb{E}\big[\tilde{X}^2\big|\tilde{L} = l\big] = \mathbb{E}\big[\tilde{X}^2\big|A_1\big]\Pr\{W^L \geq \tau\} + \mathbb{E}\big[\tilde{X}^2\big|A_2\big]\Pr\{W^L < \tau\}$$
$$= \mathbb{E}\big[(W^X)^2\mathbf{I}\{W^L \geq \tau\}\big]\mathbb{E}\big[X^2\big|L = l\big] + \mathbb{E}\big[(1 - W^X)^2\mathbf{I}\{W^L \geq \tau\}\big]\mathbb{E}\big[X^2\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\mathbf{I}\{W^L \geq \tau\}\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big]$$
$$+ \mathbb{E}\big[(W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big] + \mathbb{E}\big[(1 - W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big|L = l\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X\big]\mathbb{E}\big[X\big|L = l\big]$$
$$= \mathbb{E}\big[(W^X)^2\mathbf{I}\{W^L \geq \tau\} + (1 - W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big|L = l\big]$$
$$+ \mathbb{E}\big[(1 - W^X)^2\mathbf{I}\{W^L \geq \tau\} + (W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big]$$
$$+ 2\mathbb{E}\big[W^X(1 - W^X)\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big]. \tag{25}$$

Note that $u(W^X, W^L, \tau) = \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\}\big]$ by the definition in (8).

Now, we introduce the new random variable $\tilde{W}$ where

$$\tilde{W} = \begin{cases} 1 - W^X & \text{if } W^L \geq \tau \\ W^X & \text{if } W^L < \tau \end{cases} \tag{26}$$
$$= (1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\},$$

which implies $\mathbb{E}[\tilde{W}] = u(W^X, W^L, \tau)$ and $\mathbb{E}[1 - \tilde{W}] = 1 - u(W^X, W^L, \tau) = \mathbb{E}\big[W^X\mathbf{I}\{W^L \geq \tau\} + (1 - W^X)\mathbf{I}\{W^L < \tau\}\big]$.

If the assumption $\mathbb{E}\big[(W^X)^2\big] = \mathbb{E}\big[W^X\big]$, which is the exactly same condition of (5) in Lemma A.3, holds, then $\mathbb{E}\big[\tilde{W}^2\big] = \mathbb{E}\big[\tilde{W}\big]$ from

$$\mathbb{E}\big[\tilde{W}^2\big] = \mathbb{E}\big[\big((1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\}\big)^2\big]$$
$$= \mathbb{E}\big[(1 - W^X)^2\mathbf{I}\{W^L \geq \tau\} + (W^X)^2\mathbf{I}\{W^L < \tau\} + 0\big]$$
$$= \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X\mathbf{I}\{W^L < \tau\}\big]$$
$$= \mathbb{E}\big[\tilde{W}\big].$$

Therefore, $\mathrm{Var}\big[\tilde{W}\big] = \mathbb{E}\big[\tilde{W}^2\big] - \mathbb{E}\big[\tilde{W}\big]^2 = \mathbb{E}\big[1 - \tilde{W}\big]\mathbb{E}\big[\tilde{W}\big]$.

By using the random variable $\tilde{W}$ defined in (26), the conditioned second moment of (25) is simplified as

$$
\begin{aligned}
\mathbb{E}\big[\tilde{X}^2\big|\tilde{L} = l\big] &= \mathbb{E}\big[(W^X)^2\mathbf{I}\{W^L \geq \tau\} + (1 - W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big|L = l\big] \\
&\quad + \mathbb{E}\big[(1 - W^X)^2\mathbf{I}\{W^L \geq \tau\} + (W^X)^2\mathbf{I}\{W^L < \tau\}\big]\mathbb{E}\big[X^2\big] \\
&\quad + 2\mathbb{E}\big[W^X(1 - W^X)\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big] \\
&= \mathbb{E}\big[(1 - \tilde{W})^2\big]\mathbb{E}\big[X^2\big|L = l\big] + \mathbb{E}\big[\tilde{W}^2\big]\mathbb{E}\big[X^2\big] \\
&\quad + 2\mathbb{E}\big[\tilde{W}(1 - \tilde{W})\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big].
\end{aligned}
$$

Similarly, from (10) in Theorem A.6, the conditional mean is also simplified as

$$
\begin{aligned}
\mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big]^2 &= (1 - u(W^X, W^L, \tau))^2\mathbb{E}\big[X\big|L = l\big]^2 + u\big(W^X, W^L, \tau\big)^2\mathbb{E}\big[X\big]^2 \\
&\quad + 2(1 - u(W^X, W^L, \tau))u\big(W^X, W^L, \tau\big)\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big] \\
&= \mathbb{E}\big[1 - \tilde{W}\big]^2\mathbb{E}\big[X\big|L = l\big]^2 + \mathbb{E}\big[\tilde{W}\big]^2\mathbb{E}\big[X\big]^2 \\
&\quad + 2\mathbb{E}\big[1 - \tilde{W}\big]\mathbb{E}\big[\tilde{W}\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big].
\end{aligned}
$$

As a result, the conditional variance is determined as

$$
\begin{aligned}
\mathrm{Var}\big[\tilde{X}\big|\tilde{L} = l\big] &= \mathbb{E}\big[\tilde{X}^2\big|\tilde{L} = l\big] - \mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big]^2 \\
&= \mathbb{E}\big[(1 - \tilde{W})^2\big]\mathbb{E}\big[X^2\big|L = l\big] + \mathbb{E}\big[\tilde{W}^2\big]\mathbb{E}\big[X^2\big] + 2\mathbb{E}\big[\tilde{W}(1 - \tilde{W})\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big] \\
&\quad - \mathbb{E}\big[1 - \tilde{W}\big]^2\mathbb{E}\big[X\big|L = l\big]^2 - \mathbb{E}\big[\tilde{W}\big]^2\mathbb{E}\big[X\big]^2 - 2\mathbb{E}\big[1 - \tilde{W}\big]\mathbb{E}\big[\tilde{W}\big]\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big] \\
&= \mathbb{E}\big[(1 - \tilde{W})^2\big]\mathrm{Var}\big[X\big|L = l\big] + \mathbb{E}\big[\tilde{W}^2\big]\mathrm{Var}\big[X\big] \\
&\quad + \big(\mathbb{E}\big[(1 - \tilde{W})^2\big] - \mathbb{E}\big[1 - \tilde{W}\big]^2\big)\mathbb{E}\big[X\big|L = l\big]^2 + \big(\mathbb{E}\big[\tilde{W}^2\big] - \mathbb{E}\big[\tilde{W}\big]^2\big)\mathbb{E}\big[X\big]^2 \\
&\quad - 2\big(\mathbb{E}\big[\tilde{W}^2\big] - \mathbb{E}\big[\tilde{W}\big]^2\big)\mathbb{E}\big[X\big|L = l\big]\mathbb{E}\big[X\big] \\
&= \mathbb{E}\big[(1 - \tilde{W})^2\big]\mathrm{Var}\big[X\big|L = l\big] + \mathbb{E}\big[\tilde{W}^2\big]\mathrm{Var}\big[X\big] + \mathrm{Var}\big[\tilde{W}\big]\big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2 \\
&= \mathbb{E}\big[1 - \tilde{W}\big]\mathrm{Var}\big[X\big|L = l\big] + \mathbb{E}\big[\tilde{W}\big]\mathrm{Var}\big[X\big] + \mathbb{E}\big[1 - \tilde{W}\big]\mathbb{E}\big[\tilde{W}\big]\big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2 \\
&= \big(1 - u(W^X, W^L, \tau)\big)\mathrm{Var}\big[X\big|L = l\big] + u(W^X, W^L, \tau)\mathrm{Var}\big[X\big] \\
&\quad + \big(1 - u(W^X, W^L, \tau)\big)u(W^X, W^L, \tau)\big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2 \\
&= \mathrm{Var}\big[X\big|L = l\big] + u(W^X, W^L, \tau)\big(\mathrm{Var}\big[X\big] - \mathrm{Var}\big[X\big|L = l\big]\big) \\
&\quad + \big(1 - u(W^X, W^L, \tau)\big)u(W^X, W^L, \tau)\big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2,
\end{aligned}
$$

which is equal to

$$
\begin{aligned}
\big|\mathrm{Var}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathrm{Var}\big[X\big|L = l\big]\big| &\leq \big|u\big(W^X, W^L, \tau\big)\big| \cdot \big|\mathrm{Var}\big[X\big|L = l\big] - \mathrm{Var}\big[X\big]\big| \\
&\quad + \big|u\big(W^X, W^L, \tau\big)(1 - u\big(W^X, W^L, \tau\big))\big| \cdot \big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2.
\end{aligned}
$$

by the triangular inequality. $\qquad\square$

Lemma A.9. *Under the standard mixup scheme, $u(W, \tau) \in [0, 1]$ holds for any $\tau \in \mathbb{R}$ if $\mathbb{E}[W^2] = \mathbb{E}[W]$.*

Proof. From the definition of function $u(W, \tau)$ in (9),

$$
\begin{aligned}
u(W, \tau) &= \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq \tau\} + W\mathbf{I}\{W < \tau\}\big] \\
&= \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq \tau\} + W\big(1 - \mathbf{I}\{W \geq \tau\}\big)\big] \\
&= \mathbb{E}\big[(W)^2\big] + \mathbb{E}\big[(1 - 2W)\mathbf{I}\{W \geq \tau\}\big],
\end{aligned}
$$

where the last equality holds if $\mathbb{E}\big[(W)^2\big] = \mathbb{E}[W]$. Then, it is to trivial to show the non-negativity of $u(W, \tau)$ for all $c \in \mathbb{R}$ as follows:

$$
\begin{aligned}
u(W, \tau) &= \mathbb{E}\big[(W)^2\mathbf{I}\{W < \tau\}\big] + \mathbb{E}\big[\big(1 - 2W + (W)^2\big)\mathbf{I}\{W \geq \tau\}\big] \\
&= \mathbb{E}\big[(W)^2\mathbf{I}\{W < \tau\}\big] + \mathbb{E}\big[(1 - W)^2\mathbf{I}\{W \geq \tau\}\big] \\
&\geq 0.
\end{aligned}
$$

In a similar manner we can show the upper bound as

$$
\begin{aligned}
u(W, \tau) &= \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq \tau\} + W\mathbf{I}\{W < \tau\}\big] \\
&= \mathbb{E}\big[(1 - W)\big(1 - \mathbf{I}\{W < \tau\}\big) + W\mathbf{I}\{W < \tau\}\big] \\
&= 1 - \mathbb{E}\big[(W)^2\big] + \mathbb{E}\big[(2W - 1)\mathbf{I}\{W < \tau\}\big] \\
&= 1 - \mathbb{E}\big[(W)^2\mathbf{I}\{W \geq \tau\}\big] - \mathbb{E}\big[(1 - W)^2\mathbf{I}\{W < \tau\}\big] \\
&\leq 1.
\end{aligned}
$$

□

Lemma A.10 (Optimal Cut Point τ). *Under the standard mixup scheme with $\mathbb{E}[W^2] = \mathbb{E}[W]$, the optimal cut point τ is 0.5. That is*

$$
0.5 = \arg\min_{\tau \in \mathbb{R}} \big|u(W, \tau)\big|.
$$

Proof. From Lemma A.9, $u(W, \tau) \in [0, 1]$ holds for any $\tau \in \mathbb{R}$. Then,

$$
\begin{aligned}
\big|u(W, \tau)\big| &\geq u(W, \tau) \\
&= \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq \tau\} + W\mathbf{I}\{W < \tau\}\big] \\
&\geq \mathbb{E}\big[\min\{1 - W, W\}\mathbf{I}\{W \geq \tau\} + \min\{1 - W, W\}\mathbf{I}\{W < \tau\}\big] \\
&= \mathbb{E}\big[\min\{1 - W, W\}\big] \\
&= \mathbb{E}\big[(1 - W)\mathbf{I}\{W \geq 0.5\} + W\mathbf{I}\{W < 0.5\}\big] \\
&= u(W, 0.5)
\end{aligned}
$$

for any W. Therefore, with the condition $u(W, 0.5) \geq 0$, the optimal cut point τ for the categorical variable is 0.5. That is

$$
0.5 = \arg\min_{\tau \in (0,1)} \big|u(W, \tau)\big|.
$$

□

A.3 Proofs for Sec. 4

Corollary A.11 (Variance-Reduction mixup)**.** *For any synthetic variable $\tilde{X}$ generated by the mixup from a continuous X in (1), let the support of mixup weight variable W^X be bounded in $[0, 1]$. Then*

$$\mathrm{Var}\big[\tilde{X}\big] \leq \mathrm{Var}\big[X\big],$$

where the equality holds when $\mathrm{Pr}\big\{W^X \in \{0, 1\}\big\} = 1$.

Proof. Note that $w^2 - w \leq 0$ holds for any $w \in [0, 1]$, which implies $\mathbb{E}\big[(W^X)^2 - W^X\big] \leq 0$, where the equality condition is $\mathrm{Pr}\big\{W^X \in \{0, 1\}\big\} = 1$. From Lemma A.3, we have $\mathrm{Var}\big[\tilde{X}\big] \leq \mathrm{Var}\big[X\big]$. $\square$

Example. *Let the mixup weights are generated from the Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ where $\sigma = \sqrt{\mu - \mu^2}$ for some $\mu \in [0, 1]$, i.e., $W^X = W^Y \sim \mathrm{N}(\mu, \mu - \mu^2)$. Then, under the standard mixup scheme, we have, for any pair (X, Y), $\mathrm{Var}\big[\tilde{X}\big] = \mathrm{Var}\big[X\big]$, $\mathrm{Var}\big[\tilde{Y}\big] = \mathrm{Var}\big[Y\big]$, and $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Cov}\big[X, Y\big]$.*

Proof. By the definition of distribution $\mathrm{N}(\mu, \sigma^2)$ where $\sigma = \sqrt{\mu - \mu^2}$ for some $\mu \in [0, 1]$, $\mu = \mathbb{E}\big[(W^X)^2\big] = \mathbb{E}\big[W^X\big] = \mathbb{E}\big[(W^Y)^2\big] = \mathbb{E}\big[W^Y\big] = \mathbb{E}\big[W^X W^Y\big]$. Then, the conditions of (5) in Lemma A.3 and (6) in Theorem A.4 hold, which implies that $\mathrm{Var}\big[\tilde{X}\big] = \mathrm{Var}\big[X\big]$, $\mathrm{Var}\big[\tilde{Y}\big] = \mathrm{Var}\big[Y\big]$, and $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Cov}\big[X, Y\big]$. $\square$

Theorem A.12. *For given $\epsilon_0, \epsilon_1 \in [0, \infty)$, consider an arbitrary synthetic pair $(\tilde{X}, \tilde{Y})$ generated from (X, Y) using the standard mixup scheme with $W \sim \mathrm{EpBeta}(\alpha, \beta; \epsilon_0, \epsilon_1)$, such that $\alpha, \beta \in (0, \infty)$ satisfy*

$$(1 + \epsilon_1 - \epsilon_0(\beta/\alpha)) \cdot (1 + \epsilon_0 - \epsilon_1(\alpha/\beta)) \cdot (1 + \alpha + \beta) = (1 + \epsilon_0 + \epsilon_1)^2.$$

Then we have $\mathrm{Var}\big[\tilde{X}\big] = \mathrm{Var}\big[X\big]$ and $\mathrm{Cov}\big[\tilde{X}, \tilde{Y}\big] = \mathrm{Cov}\big[X, Y\big]$.

Proof. Let W follow the $\mathrm{Beta}(\alpha, \beta)$ distribution, which implies $\mathbb{E}[W] = \frac{\alpha}{\alpha+\beta}$ and $\mathrm{Var}[W] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Then,

$$\begin{aligned}
\mathbb{E}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big] &= \frac{(1 + \epsilon_0 + \epsilon_1)\alpha}{\alpha + \beta} - \epsilon_0 \\
&= \frac{(1 + \epsilon_1)\alpha - \epsilon_0\beta}{\alpha + \beta}, \\
\mathbb{E}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big]^2 &= \mathrm{Var}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big] + \big(\mathbb{E}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big]\big)^2 \\
&= \frac{(1 + \epsilon_0 + \epsilon_1)^2\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \left(\frac{(1 + \epsilon_1)\alpha - \epsilon_0\beta}{\alpha + \beta}\right)^2.
\end{aligned}$$

By Theorem A.4, $\mathrm{Var}\big[\tilde{D}\big] = \mathrm{Var}\big[D\big]$ if and only if

$$\mathbb{E}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big] = \mathbb{E}\big[(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0\big]^2. \tag{27}$$

To find the $\alpha, \beta \in (0, \infty)$ that satisfy (27),

$$((1 + \epsilon_1)\alpha - \epsilon_0\beta)(\alpha + \beta) = \frac{(1 + \epsilon_0 + \epsilon_1)^2\alpha\beta}{(\alpha + \beta + 1)} + \big((1 + \epsilon_1)\alpha - \epsilon_0\beta\big)^2,$$

which is equal to

$$(1 + \epsilon_1 - \epsilon_0(\beta/\alpha)) \cdot (1 + \epsilon_0 - \epsilon_1(\alpha/\beta)) \cdot (1 + \alpha + \beta) = (1 + \epsilon_0 + \epsilon_1)^2.$$

Note that the above equation can be rewritten as

$$\alpha + \beta = \frac{(1 + \epsilon_0 + \epsilon_1)^2}{(1 + \epsilon_1 - \epsilon_0(\beta/\alpha)) \cdot (1 + \epsilon_0 - \epsilon_1(\alpha/\beta))} - 1,$$

which implies that we can find (α, β) that satisfies the condition in (27) for given $\beta/\alpha, \epsilon_0$, and ϵ_1. $\square$

Theorem A.13. *Consider an arbitrary synthetic triple $(\tilde{X}, \tilde{Y}, \tilde{L})$ generated from (X, Y, L) using the standard mixup scheme with $W \sim \mathrm{EpBeta}(\alpha, \beta; \epsilon_0, \epsilon_1)$ for given $\epsilon_0, \epsilon_1 \in [0, \infty)$ and $\tau = 0.5$. Now suppose that, for a given $\delta \in [0, 1]$, (α, β) satisfies (17) and the following*

$$\frac{1 + \epsilon_0 - \epsilon_1 \alpha/\beta}{1 + \alpha/\beta} + \frac{2(1 + \epsilon_0 + \epsilon_1)}{1 + \beta/\alpha} \frac{B(\tilde{\epsilon}; \alpha + 1, \beta)}{B(1; \alpha + 1, \beta)} - (1 + 2\epsilon_0) \frac{B(\tilde{\epsilon}; \alpha, \beta)}{B(1; \alpha, \beta)} \le \delta,$$

where $B(x; \alpha, \beta) = \int_0^b t^{\alpha - 1}(1 - t)^{\beta - 1} dt$ is the incomplete beta function and $\tilde{\epsilon} = \frac{0.5 + \epsilon_0}{1 + \epsilon_0 + \epsilon_1}$.

Then, the gap of conditional (on categorical L) mean and variance are bounded as follows:

$$\left| \mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathbb{E}\big[X\big|L = l\big] \right| = \delta \cdot \Pr\{L \ne l\} \cdot \left| \mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big|L \ne l\big] \right|$$

and

$$\left| \mathrm{Var}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathrm{Var}\big[X\big|L = l\big] \right| \le \delta \cdot \left| \mathrm{Var}\big[X\big|L = l\big] - \mathrm{Var}\big[X\big] \right| + \delta(1 - \delta) \cdot \big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2.$$

Proof. Note that the conditional mean and variance gaps are bounded by the functions of $u(W^X, W^L, \tau)$ from Corollary A.7 and Theorem A.8, respectively. To determine it, define the random variable $W \sim \mathrm{Beta}(\alpha, \beta)$, which implies $W^X = W^L = (1 + \epsilon_0 + \epsilon_1)W - \epsilon_0$. Then, with (18),

$$\begin{aligned}
u\big(W^X, W^L, \tau\big) &= u(W^X, 0.5) \\
&= \mathbb{E}\big[(1 - W^X)\mathbf{I}\{W^X \ge 0.5\} + W^X\mathbf{I}\{W^X < 0.5\}\big] \\
&= \mathbb{E}\big[(1 - W^X)(1 - \mathbf{I}\{W^X < 0.5\}) + W^X\mathbf{I}\{W^X < 0.5\}\big] \\
&= \mathbb{E}\big[1 - W^X\big] + \mathbb{E}\big[(2W^X - 1)\mathbf{I}\{W^X < 0.5\}\big] \\
&= \frac{-\epsilon_1\alpha + (1 + \epsilon_0)\beta}{\alpha + \beta} + \mathbb{E}\big[(2(1 + \epsilon_0 + \epsilon_1)W - 1 - 2\epsilon_0)\mathbf{I}\{(1 + \epsilon_0 + \epsilon_1)W - \epsilon_0 < 0.5\}\big] \\
&= \frac{-\epsilon_1\alpha + (1 + \epsilon_0)\beta}{\alpha + \beta} + \mathbb{E}\left[\left(2(1 + \epsilon_0 + \epsilon_1)W - 1 - 2\epsilon_0\right)\mathbf{I}\left\{W < \frac{0.5 + \epsilon_0}{1 + \epsilon_0 + \epsilon_1}\right\}\right] \\
&= \frac{1 + \epsilon_0 - \epsilon_1\alpha/\beta}{1 + \alpha/\beta} + 2(1 + \epsilon_0 + \epsilon_1)\frac{B(\tilde{\epsilon}; \alpha + 1, \beta)}{B(1; \alpha, \beta)} - (1 + 2\epsilon_0)\frac{B(\tilde{\epsilon}; \alpha, \beta)}{B(1; \alpha, \beta)} \\
&= \frac{1 + \epsilon_0 - \epsilon_1\alpha/\beta}{1 + \alpha/\beta} + 2(1 + \epsilon_0 + \epsilon_1)\frac{\alpha}{\alpha + \beta}\frac{B(\tilde{\epsilon}; \alpha + 1, \beta)}{B(1; \alpha + 1, \beta)} - (1 + 2\epsilon_0)\frac{B(\tilde{\epsilon}; \alpha, \beta)}{B(1; \alpha, \beta)} \\
&= \frac{1 + \epsilon_0 - \epsilon_1\alpha/\beta}{1 + \alpha/\beta} + \frac{2(1 + \epsilon_0 + \epsilon_1)}{1 + \beta/\alpha}\frac{B(\tilde{\epsilon}; \alpha + 1, \beta)}{B(1; \alpha + 1, \beta)} - (1 + 2\epsilon_0)\frac{B(\tilde{\epsilon}; \alpha, \beta)}{B(1; \alpha, \beta)} \\
&\le \delta,
\end{aligned}$$

where $B(x; \alpha, \beta) = \int_0^b t^{\alpha - 1}(1 - t)^{\beta - 1} dt$ is the incomplete beta function, and $\tilde{\epsilon} = \frac{0.5 + \epsilon_0}{1 + \epsilon_0 + \epsilon_1}$.

Note that $u(W^X, \tau)$ is bounded in $[0, 1]$ under the condition of (17) by Lemma A.9, which implies $\left|u(W^X, W^L, \tau)\right| \le \delta$. From Corollary A.7 and Theorem A.8, respectively,

$$\begin{aligned}
\left| \mathbb{E}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathbb{E}\big[X\big|L = l\big] \right| &= \left|u\big(W^X, W^L, \tau\big)\right| \cdot \Pr\{L \ne l\} \cdot \left| \mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big|L \ne l\big] \right| \\
&\le \delta \cdot \Pr\{L \ne l\} \cdot \left| \mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big|L \ne l\big] \right|, \\
\left| \mathrm{Var}\big[\tilde{X}\big|\tilde{L} = l\big] - \mathrm{Var}\big[X\big|L = l\big] \right| &\le \left|u\big(W^X, W^L, \tau\big)\right| \cdot \left| \mathrm{Var}\big[X\big|L = l\big] - \mathrm{Var}\big[X\big] \right| \\
&\quad + \left|u\big(W^X, W^L, \tau\big)\big(1 - u\big(W^X, W^L, \tau\big)\big)\right| \cdot \big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2 \\
&\le \delta \cdot \left| \mathrm{Var}\big[X\big|L = l\big] - \mathrm{Var}\big[X\big] \right| + \delta(1 - \delta) \cdot \big(\mathbb{E}\big[X\big|L = l\big] - \mathbb{E}\big[X\big]\big)^2. \quad (28)
\end{aligned}$$

The last inequality in (28) comes from

$$\left|u\big(W^X, W^L, \tau\big)\big(1 - u\big(W^X, W^L, \tau\big)\big)\right| = u\big(W^X, W^L, \tau\big) - u\big(W^X, W^L, \tau\big)^2 \le u\big(W^X, W^L, \tau\big) \le \delta.$$

Therefore, the gap of conditional mean and variance are bounded as

$$\left|\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big]-\mathbb{E}\big[X\big|L=l\big]\right|=\delta\cdot\Pr\{L\neq l\}\cdot\left|\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big|L\neq l\big]\right|,$$

and

$$\left|\mathrm{Var}\big[\tilde{X}\big|\tilde{L}=l\big]-\mathrm{Var}\big[X\big|L=l\big]\right|\leq\delta\cdot\left|\mathrm{Var}\big[X\big|L=l\big]-\mathrm{Var}\big[X\big]\right|+\delta(1-\delta)\cdot\big(\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big]\big)^2.$$

Moreover, note that $\Pr\{L\neq l\}$, $\left|\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big|L\neq l\big]\right|$, $\left|\mathrm{Var}\big[X\big|L=l\big]-\mathrm{Var}\big[X\big]\right|$, and $\big(\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big]\big)$ are constants given by the original distribution $\mathfrak{D}$. As result, when δ goes to 0,

$$\left|\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big]-\mathbb{E}\big[X\big|L=l\big]\right|\leq\delta\cdot\Pr\{L\neq l\}\cdot\left|\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big|L\neq l\big]\right|$$
$$\to 0,$$
$$\left|\mathrm{Var}\big[\tilde{X}\big|\tilde{L}=l\big]-\mathrm{Var}\big[X\big|L=l\big]\right|\leq\delta\cdot\left|\mathrm{Var}\big[X\big|L=l\big]-\mathrm{Var}\big[X\big]\right|+\delta(1-\delta)\cdot\big(\mathbb{E}\big[X\big|L=l\big]-\mathbb{E}\big[X\big]\big)^2$$
$$\to 0,$$

which are equal to

$$\mathbb{E}\big[\tilde{X}\big|\tilde{L}=l\big]\to\mathbb{E}\big[X\big|L=l\big],$$
$$\mathrm{Var}\big[\tilde{X}\big|\tilde{L}=l\big]\to\mathrm{Var}\big[X\big|L=l\big],$$

as δ goes to 0. $\qquad\square$

B EXPERIMENT DETAILS

B.1 Data Descriptions and Synthesis Details

In this experiment, we select 6 popular datasets used in Gorishniy et al. (2021); Kotelnikov et al. (2023). For convenience purposes, the instances with missing values are removed. The preprocessed datasets are summarized in Table 2.

Table 2: Data description.

Name	# Instance	# Num	# Cat	Task	License	Source
Abalone	4177	8	1	Regress	CC4.0	UCI ML (Nash et al., 1995)
CA Housing	20433	9	1	Regress	CC0	Kaggle (Géron, 2022)
House 16H	22784	17	0	Regress	Public	OpenML
Adult	48842	6	9	Classify	CC4.0	UCI ML (Becker and Kohavi, 1996)
Diabetes	768	8	1	Classify	Public	OpenML
Wilt	4839	5	1	Classify	Public	OpenML (Johnson et al., 2013)

We synthesize data using the mixup method with various weight distributions, such as EpBeta, Beta, and Unif, and compare them against four baseline methods implemented through an open-source code (Qian et al., 2023); TVAE, CTGAN (Xu et al., 2019), TabDDPM (Kotelnikov et al., 2023), and GReaT (Borisov et al., 2023). Each method is used with its default settings. For example, while TabDDPM allows for the selection of the target variable as an input, we do not utilize this option.

We apply the EpBeta distribution with four different δ values: $\delta \in 0.001, 0.005, 0.01, 0.05$, as well as the Beta$(0.1, 0.1)$ and Unif$(0, 1)$ distributions as Mixup weight distributions, resulting in 10 synthetic datasets for each original dataset.

Using two NVIDIA GeForce RTX 3090 GPUs, we report the model training and generation times for producing a single synthetic dataset with the same number of instances as the original. These results are shown in Table 3. Note that for the 'House 16H' and 'Adult' datasets, the GReaT method is excluded because it fails to converge within the default number of epochs.

Table 3: Training and generating time (Seconds).

Name	Mixup	TVAE	CTGAN	TabDDPM	GReaT
Abalone	0.004	207	213	73	6001
CA Housing	0.009	827	702	362	28852
House 16H	0.016	2364	2713	239	-
Adult	0.046	4803	4533	1042	-
Diabetes	0.002	39	131	18	1277
Wilt	0.004	261	427	90	6843

B.2 Relative Bias of Synthetic Data

We compare the relative bias of covariance and expectation of continuous variables from each synthetic dataset, calculated as $\frac{\text{Cov}[\tilde{X},\tilde{Y}]-\text{Cov}[X,Y]}{\text{Cov}[X,Y]}$ for covariance and $\frac{\mathbb{E}[\tilde{X}]-\mathbb{E}[X]}{\mathbb{E}[X]}$ for expectation. In the figures, negative bias is shown in blue, positive bias in red, and grey indicates bias close to zero. Due to space constraints, we present results for only the first synthesized dataset ($m = n$) and the combined results from five subsequently synthesized datasets, equivalent to generating synthetic data with five times the number of original instances ($m = 5n$). Although some small differences in expectation and (co)variance may appear in a single dataset synthesized using the EpBeta distribution, these differences diminish as the number of synthesized instances increases across all datasets. Additionally, we abbreviate the annotation for 'House 16H' due to the large number of variables.
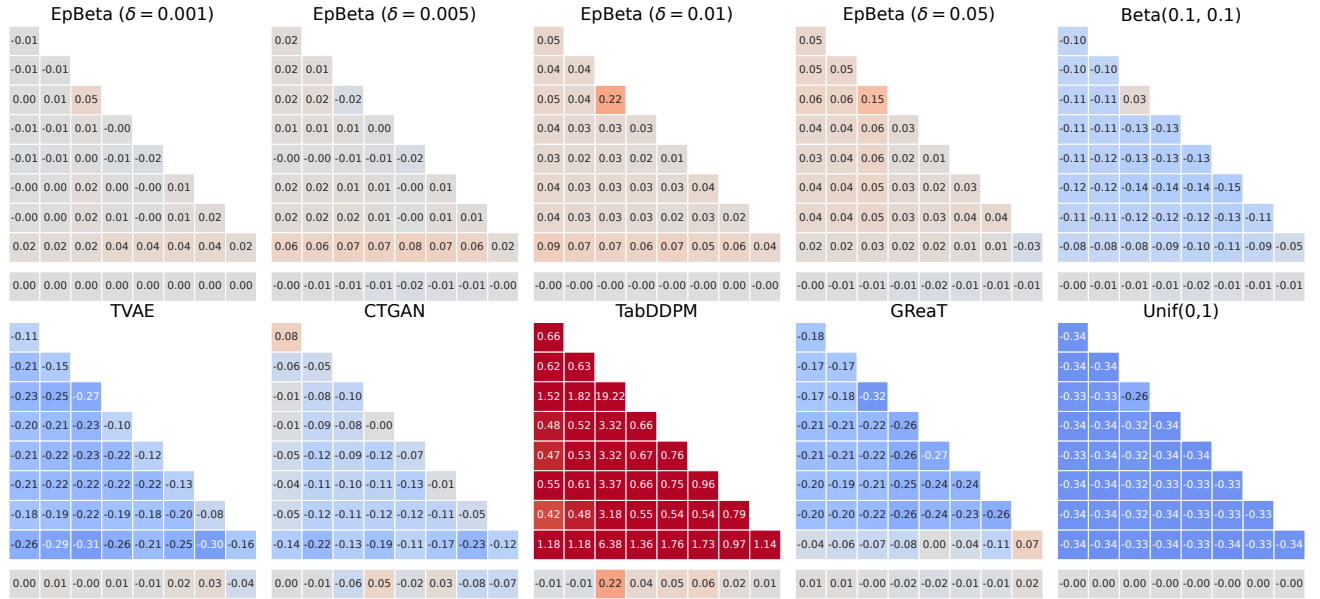
Figure 5: The relative bias of (co)variance (triangle) and expectation (bar) for 'Abalone' with $m = n$.
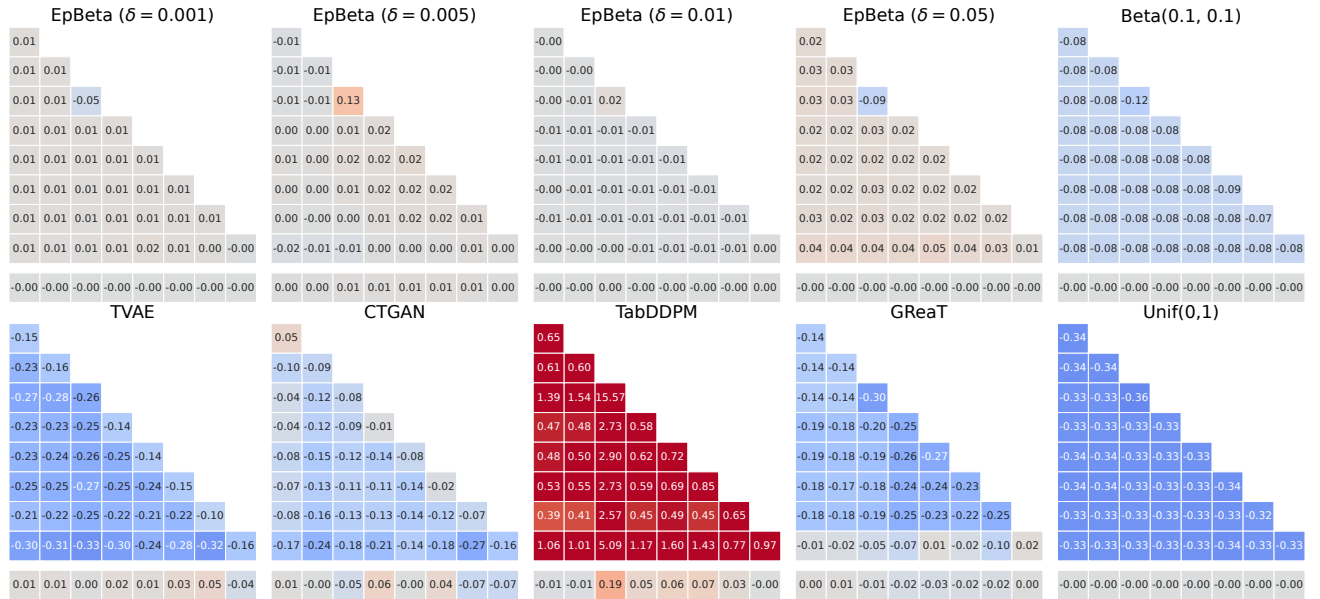

Figure 6: The relative bias of (co)variance (triangle) and expectation (bar) for 'Abalone' with $m = 5n$.
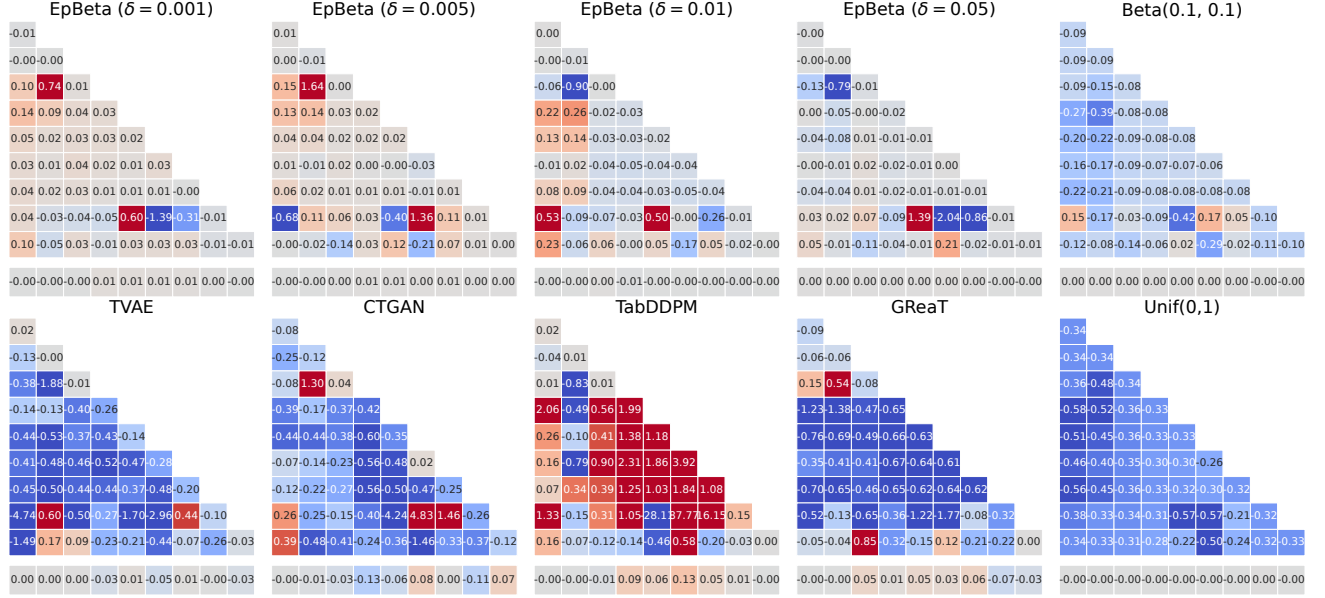
Figure 7: The relative bias of (co)variance (triangle) and expectation (bar) for 'CA Housing' with $m = n$.
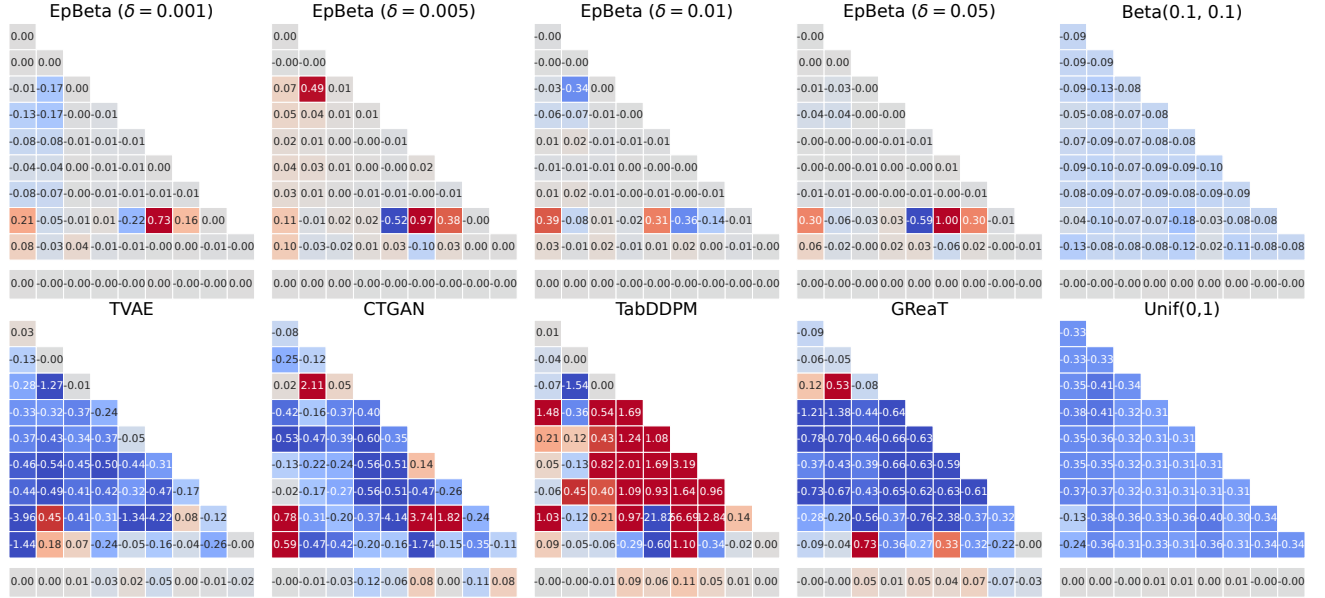

Figure 8: The relative bias of (co)variance (triangle) and expectation (bar) for 'CA Housing' with $m = 5n$.
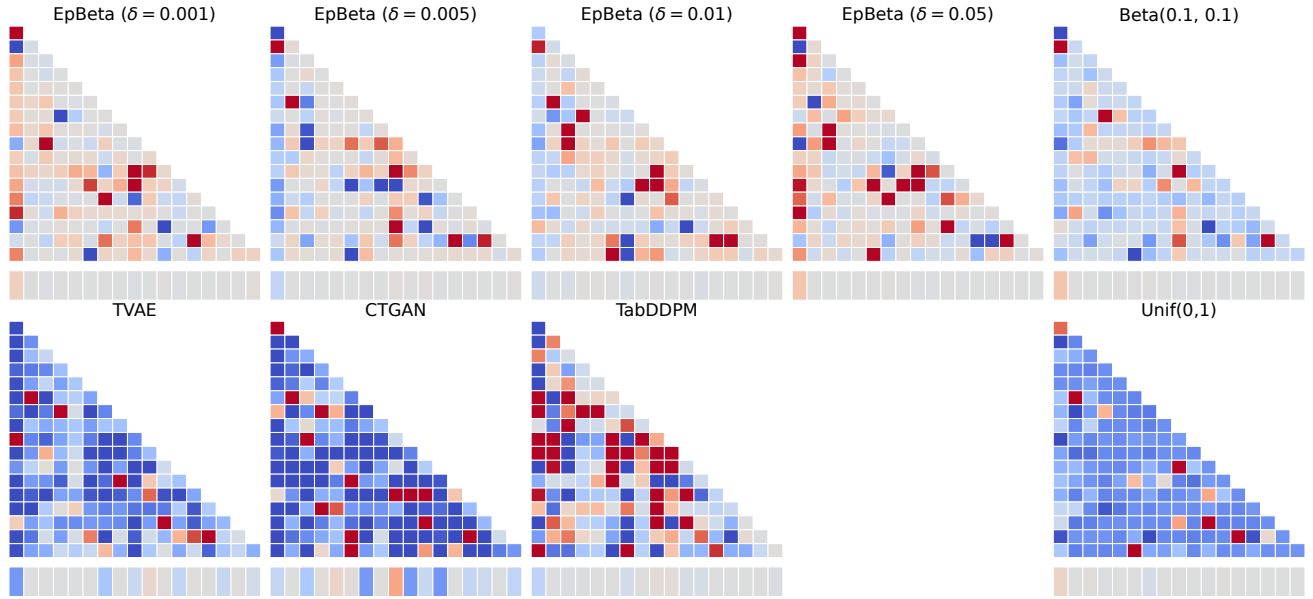
Figure 9: The relative bias of (co)variance (triangle) and expectation (bar) for 'House 16H' with $m = n$.
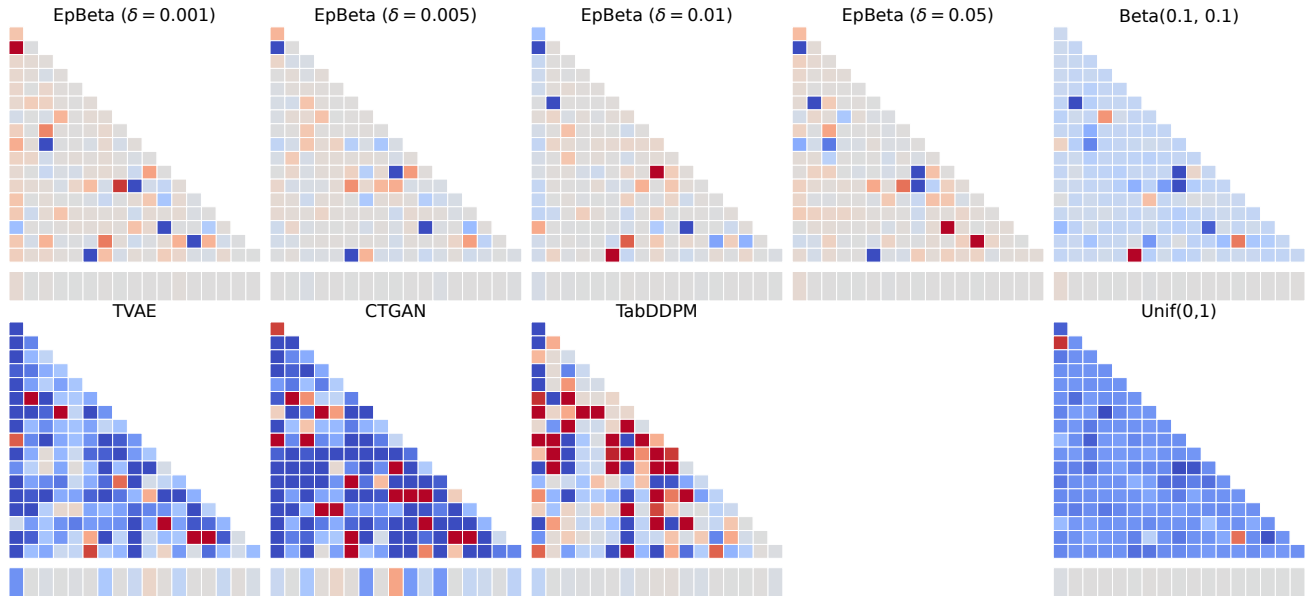

Figure 10: The relative bias of (co)variance (triangle) and expectation (bar) for 'House 16H' with $m = 5n$.

Figure 11: The relative bias of (co)variance (triangle) and expectation (bar) for 'Adult' with $m = n$.

Figure 12: The relative bias of (co)variance (triangle) and expectation (bar) for 'Adult' with $m = 5n$.

Figure 13: The relative bias of (co)variance (triangle) and expectation (bar) for 'Diabetes' with $m = n$.

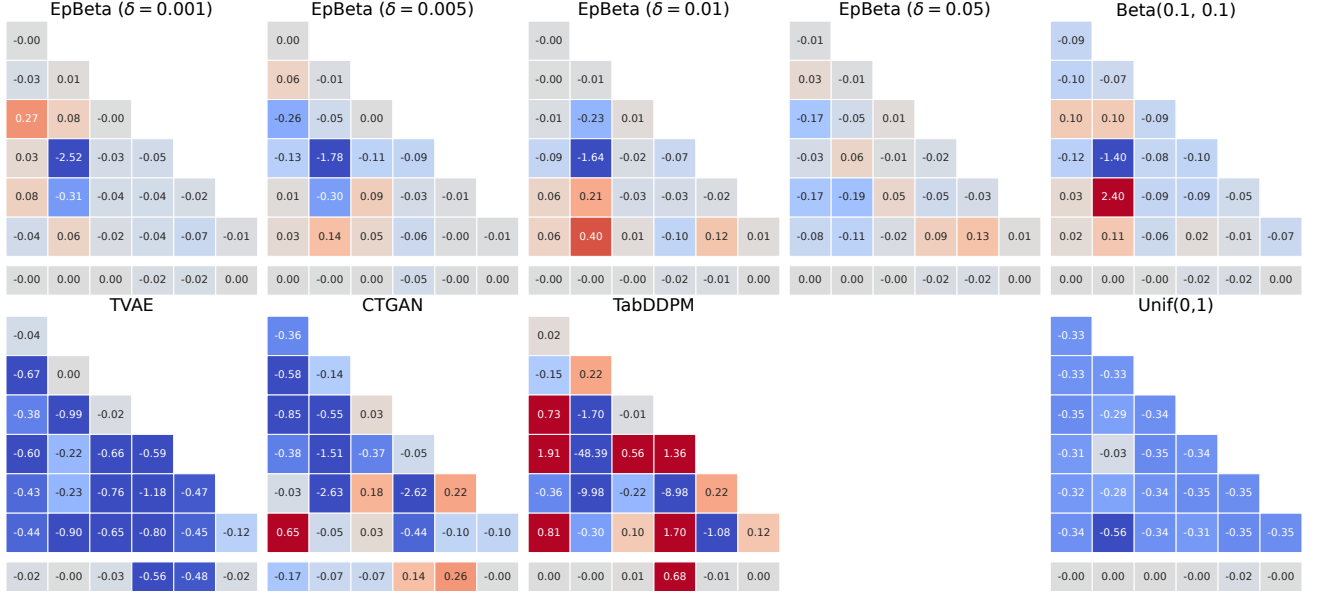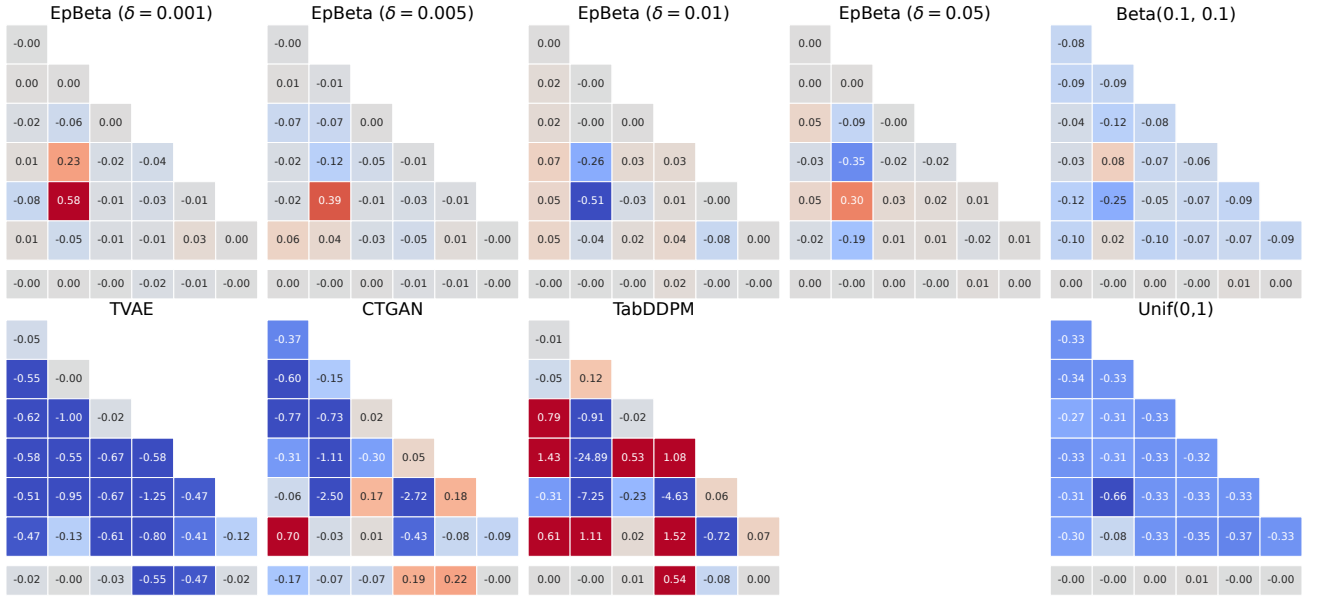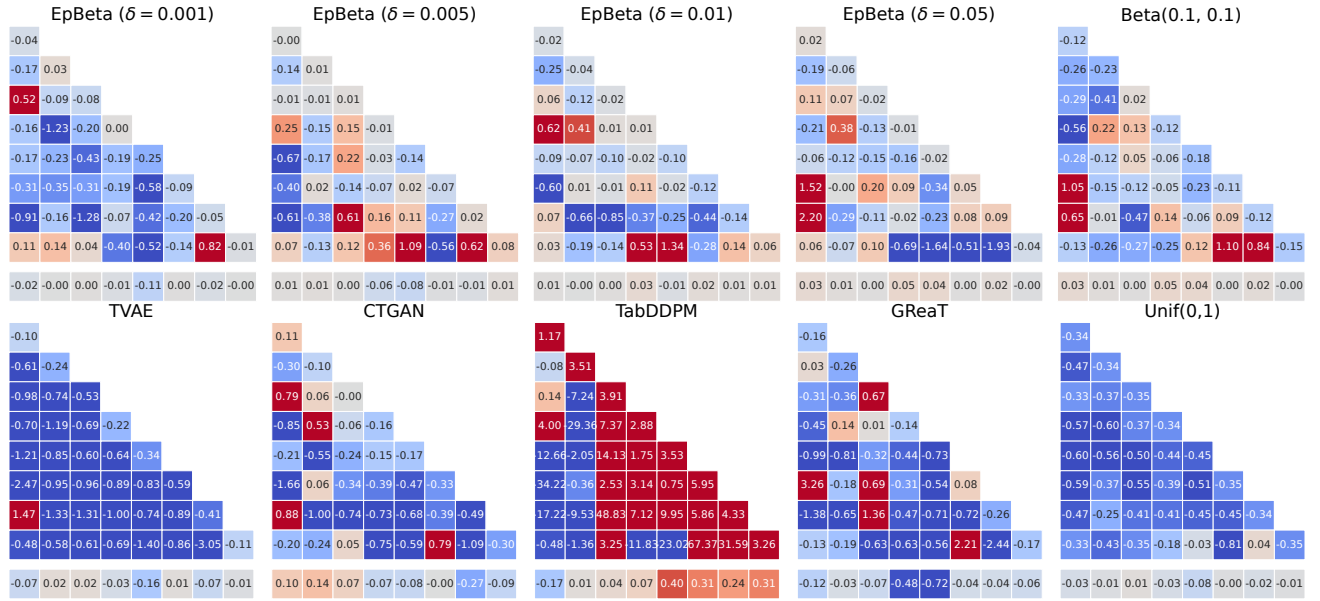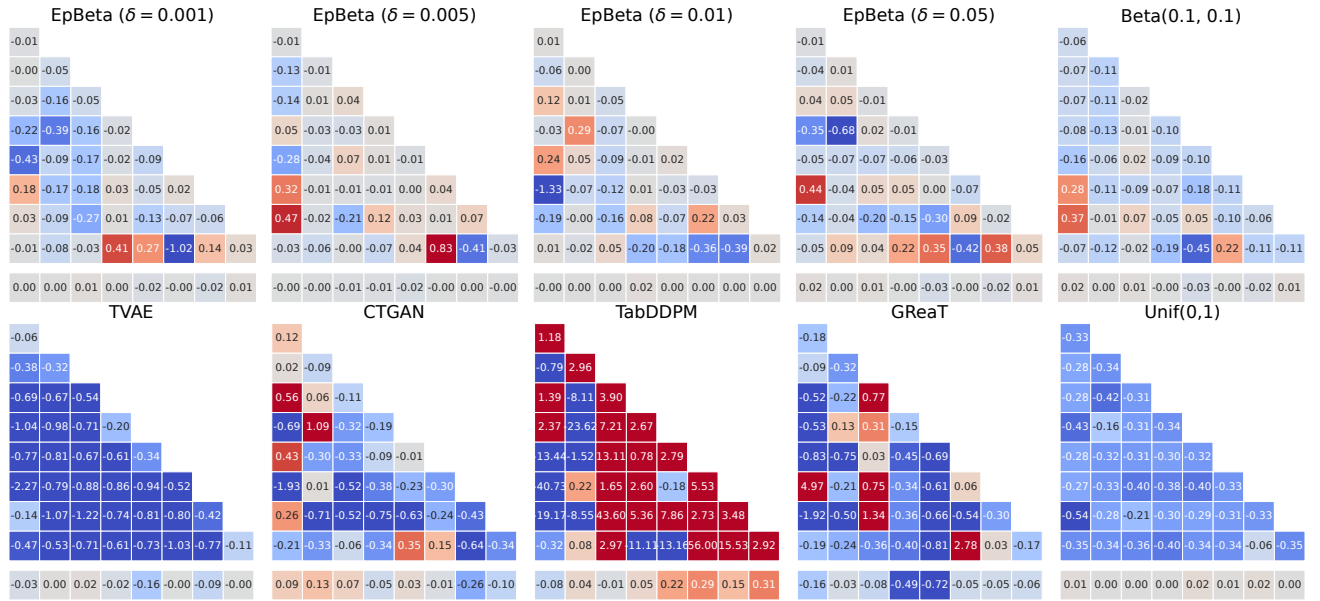Figure 14: The relative bias of (co)variance (triangle) and expectation (bar) for 'Diabetes' with $m = 5n$.
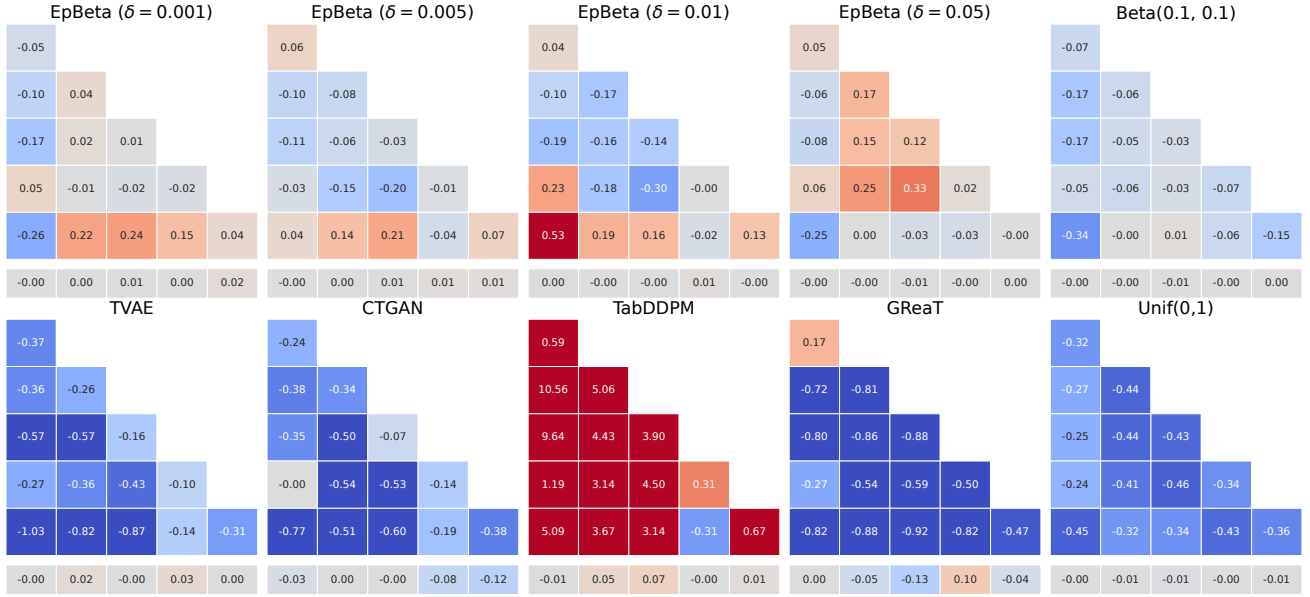
Figure 15: The relative bias of (co)variance (triangle) and expectation (bar) for 'Wilt' with $m = n$.
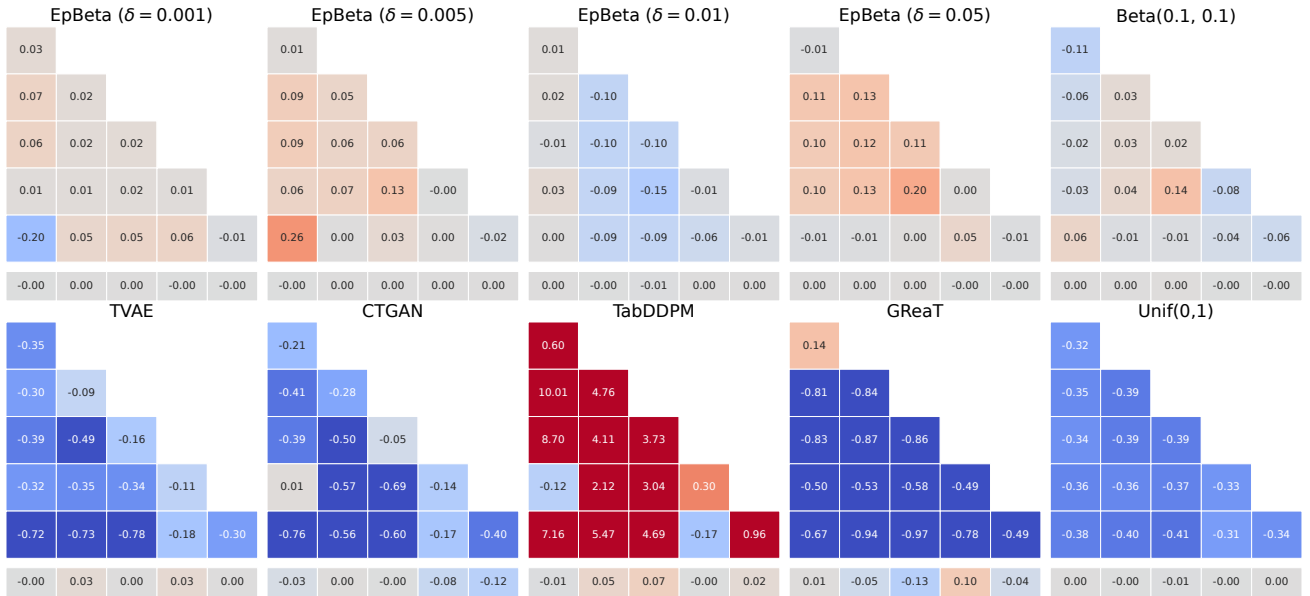

Figure 16: The relative bias of (co)variance (triangle) and expectation (bar) for 'Wilt' with $m = 5n$.

B.3 Machine Learning Efficiency

We assess machine learning efficiency by training various models on different synthetic datasets, following the experimental protocol of Gorishniy et al. (2021); Zhao et al. (2021); Kotelnikov et al. (2023). Our focus is on evaluating how closely each synthetic dataset resembles the original data, rather than on the effectiveness of the trained models using these synthetic datasets.

We select five models for training: a linear model (logistic regression for classification tasks and ridge regression for regression tasks), a decision tree, a random forest, a multilayer perceptron (MLP), and CatBoost (Prokhorenkova et al., 2018). We implement four of these models using the Scikit-learn library (Pedregosa et al., 2011), excluding CatBoost. We employ the same hyperparameters specified in Kotelnikov et al. (2023), utilizing a min-max scaler exclusively for training the MLP.

To evaluate model performance, we generate 10 sets of synthetic data for each method. We then individually train five models with different random seeds for each synthetic dataset, resulting in a total of 50 models for each synthetic method. Finally, we evaluate the R-squared value for regression tasks or accuracy for classification tasks against the original data. The table below presents the average and standard deviation of the evaluation metrics for each model. The results indicate that the performance of the mixup-driven synthetic datasets is comparable to that of other machine learning-based synthetic methods.

Table 4: The performance (R-squared or accuracy) of linear models.

Name	Abalone	CA Housing	House 16H	Adult	Diabetes	Wilt
Original	0.5355	0.6465	0.2527	0.8249	0.7812	0.9682
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
EpBeta	0.5228	0.6360	0.2517	0.8140	0.6503	0.9461
$(\delta = 0.001)$	(0.0023)	(0.0010)	(0.0018)	(0.0205)	(0.0083)	(0.0000)
EpBeta	0.5230	0.6365	0.2495	0.8123	0.6513	0.9461
$(\delta = 0.005)$	(0.0022)	(0.0004)	(0.0054)	(0.0233)	(0.0084)	(0.0000)
EpBeta	0.5237	0.6367	0.2506	0.8106	0.6522	0.9461
$(\delta = 0.01)$	(0.0023)	(0.0003)	(0.0026)	(0.0252)	(0.0025)	(0.0000)
EpBeta	0.5238	0.6364	0.2517	0.8075	0.6561	0.9461
$(\delta = 0.05)$	(0.0023)	(0.0007)	(0.0015)	(0.0243)	(0.0060)	(0.0000)
Beta	0.5247	0.6368	0.2500	0.8022	0.6527	0.9461
	(0.0027)	(0.0005)	(0.0012)	(0.0264)	(0.0032)	(0.0000)
Unif	0.5264	0.6380	0.2490	0.8159	0.6978	0.9452
	(0.0014)	(0.0005)	(0.0009)	(0.0143)	(0.0264)	(0.0003)
TVAE	0.4006	0.6009	0.1634	0.7963	0.7400	0.9461
	(0.0080)	(0.0025)	(0.0173)	(0.0118)	(0.0170)	(0.0000)
CTGAN	0.3954	0.5122	0.1338	0.8151	0.7671	0.9461
	(0.0120)	(0.0027)	(0.0038)	(0.0056)	(0.0117)	(0.0000)
TabDDPM	0.3829	0.6036	0.2216	0.8207	0.7469	0.9482
	(0.0131)	(0.0040)	(0.0110)	(0.0068)	(0.0087)	(0.0035)
GReaT	0.5156	0.6364	-	-	0.7520	0.9570
	(0.0026)	(0.0018)			(0.0093)	(0.0066)

Table 5: The performance (R-squared or accuracy) of tree models.

Name	Abalone	CA Housing	House 16H	Adult	Diabetes	Wilt
Original	1.0000	1.0000	0.9999	0.9731	1.0000	1.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)
EpBeta	0.4727	0.7334	0.5852	0.8001	0.5719	0.9014
($\delta = 0.001$)	(0.0226)	(0.0081)	(0.0106)	(0.0016)	(0.0213)	(0.0083)
EpBeta	0.3721	0.6642	0.5196	0.8010	0.5588	0.8990
($\delta = 0.005$)	(0.0297)	(0.0074)	(0.0218)	(0.0032)	(0.0169)	(0.0070)
EpBeta	0.3496	0.6327	0.4903	0.8000	0.5572	0.8999
($\delta = 0.01$)	(0.0352)	(0.0082)	(0.0306)	(0.0028)	(0.0274)	(0.0059)
EpBeta	0.2367	0.5315	0.3722	0.8052	0.5742	0.9027
($\delta = 0.05$)	(0.0308)	(0.0102)	(0.0288)	(0.0021)	(0.0286)	(0.0052)
Beta	0.5455	0.7709	0.5956	0.7952	0.5558	0.8987
	(0.0267)	(0.0081)	(0.0216)	(0.0031)	(0.0215)	(0.0055)
Unif	0.2969	0.5138	0.2477	0.7979	0.5796	0.9047
	(0.0367)	(0.0099)	(0.0388)	(0.0036)	(0.0352)	(0.0034)
TVAE	-0.0660	0.2159	-0.1702	0.7584	0.6024	0.9395
	(0.0485)	(0.0198)	(0.0528)	(0.0046)	(0.0329)	(0.0085)
CTGAN	-0.1381	-0.0740	-0.6111	0.7390	0.6295	0.9314
	(0.0528)	(0.0289)	(0.1483)	(0.0067)	(0.0265)	(0.0053)
TabDDPM	0.1679	0.4876	0.1382	0.7975	0.6910	0.9780
	(0.0369)	(0.0096)	(0.0318)	(0.0020)	(0.0152)	(0.0025)
GReaT	0.0193	0.5312	-	-	0.6826	0.9345
	(0.0455)	(0.0122)			(0.0147)	(0.0164)

Table 6: The performance (R-squared or accuracy) of random forest models.

Name	Abalone	CA Housing	House 16H	Adult	Diabetes	Wilt
Original	0.9356	0.9758	0.9496	0.9550	1.0000	1.0000
	(0.0012)	(0.0001)	(0.0007)	(0.0008)	(0.0000)	(0.0000)
EpBeta	0.7091	0.8676	0.7970	0.8404	0.6138	0.9407
($\delta = 0.001$)	(0.0063)	(0.0027)	(0.0079)	(0.0007)	(0.0169)	(0.0014)
EpBeta	0.6742	0.8403	0.7680	0.8405	0.6170	0.9429
($\delta = 0.005$)	(0.0056)	(0.0017)	(0.0082)	(0.0009)	(0.0162)	(0.0009)
EpBeta	0.6542	0.8240	0.7509	0.8409	0.6102	0.9438
($\delta = 0.01$)	(0.0055)	(0.0028)	(0.0079)	(0.0009)	(0.0177)	(0.0008)
EpBeta	0.6066	0.7751	0.6932	0.8438	0.6321	0.9451
($\delta = 0.05$)	(0.0064)	(0.0028)	(0.0047)	(0.0010)	(0.0279)	(0.0007)
Beta	0.7334	0.8758	0.7973	0.8425	0.5883	0.9247
	(0.0103)	(0.0029)	(0.0061)	(0.0011)	(0.0185)	(0.0020)
Unif	0.5866	0.7370	0.6423	0.8481	0.6567	0.9449
	(0.0061)	(0.0029)	(0.0088)	(0.0009)	(0.0257)	(0.0008)
TVAE	0.4213	0.6834	0.4450	0.8169	0.7075	0.9700
	(0.0047)	(0.0054)	(0.0102)	(0.0023)	(0.0161)	(0.0016)
CTGAN	0.4263	0.4746	0.2409	0.8336	0.7408	0.9663
	(0.0121)	(0.0107)	(0.0136)	(0.0018)	(0.0172)	(0.0022)
TabDDPM	0.5351	0.7493	0.6017	0.8484	0.7589	0.9850
	(0.0037)	(0.0018)	(0.0058)	(0.0014)	(0.0084)	(0.0013)
GReaT	0.5047	0.7809	-	-	0.7554	0.9696
	(0.0120)	(0.0033)			(0.0114)	(0.0064)

Table 7: The performance (R-squared or accuracy) of MLP models.

Name	Abalone	CA Housing	House 16H	Adult	Diabetes	Wilt
Original	0.5404	0.7324	0.4838	0.8767	0.7792	0.9461
	(0.0064)	(0.0069)	(0.0058)	(0.0006)	(0.0043)	(0.0000)
EpBeta	0.5155	0.7019	0.4564	0.8403	0.6510	0.9461
($\delta = 0.001$)	(0.0117)	(0.0126)	(0.0104)	(0.0015)	(0.0016)	(0.0000)
EpBeta	0.5119	0.6902	0.4363	0.8406	0.6518	0.9461
($\delta = 0.005$)	(0.0146)	(0.0098)	(0.0154)	(0.0019)	(0.0029)	(0.0000)
EpBeta	0.5090	0.6851	0.4298	0.8414	0.6514	0.9461
($\delta = 0.01$)	(0.0163)	(0.0093)	(0.0124)	(0.0011)	(0.0016)	(0.0000)
EpBeta	0.4979	0.6697	0.3963	0.8424	0.6511	0.9461
($\delta = 0.05$)	(0.0201)	(0.0090)	(0.0226)	(0.0016)	(0.0008)	(0.0000)
Beta	0.5201	0.7051	0.4606	0.8430	0.6513	0.9461
	(0.0121)	(0.0118)	(0.0149)	(0.0018)	(0.0018)	(0.0000)
Unif	0.5113	0.6834	0.4310	0.8500	0.6795	0.9461
	(0.0146)	(0.0085)	(0.0115)	(0.0019)	(0.0166)	(0.0000)
TVAE	0.4085	0.6341	0.3387	0.8167	0.7124	0.9461
	(0.0195)	(0.0094)	(0.0240)	(0.0059)	(0.0181)	(0.0000)
CTGAN	0.3971	0.5090	0.2411	0.8125	0.7641	0.9461
	(0.0287)	(0.0352)	(0.0145)	(0.0060)	(0.0123)	(0.0000)
TabDDPM	0.4742	0.7133	0.4478	0.8404	0.7527	0.9461
	(0.0241)	(0.0061)	(0.0120)	(0.0019)	(0.0115)	(0.0000)
GReaT	0.5086	0.7117	-	-	0.7464	0.9053
	(0.0152)	(0.0093)			(0.0099)	(0.0203)

Table 8: The performance (R-squared or accuracy) of CatBoost models.

Name	Abalone	CA Housing	House 16H	Adult	Diabetes	Wilt
Original	0.8188	0.9736	0.9618	0.8990	0.9958	1.0000
	(0.0007)	(0.0001)	(0.0001)	(0.0004)	(0.0005)	(0.0000)
EpBeta	0.6796	0.8945	0.8411	0.8368	0.6138	0.9288
($\delta = 0.001$)	(0.0070)	(0.0019)	(0.0056)	(0.0012)	(0.0224)	(0.0021)
EpBeta	0.6539	0.8713	0.8185	0.8365	0.6158	0.9326
($\delta = 0.005$)	(0.0071)	(0.0013)	(0.0059)	(0.0015)	(0.0140)	(0.0030)
EpBeta	0.6378	0.8590	0.8042	0.8363	0.6088	0.9354
($\delta = 0.01$)	(0.0055)	(0.0026)	(0.0050)	(0.0019)	(0.0218)	(0.0016)
EpBeta	0.6007	0.8216	0.7596	0.8383	0.6267	0.9401
($\delta = 0.05$)	(0.0067)	(0.0030)	(0.0047)	(0.0016)	(0.0120)	(0.0015)
Beta	0.6934	0.8976	0.8335	0.8341	0.5919	0.9249
	(0.0085)	(0.0016)	(0.0053)	(0.0016)	(0.0244)	(0.0027)
Unif	0.5862	0.7996	0.7189	0.8400	0.6355	0.9410
	(0.0053)	(0.0030)	(0.0080)	(0.0018)	(0.0209)	(0.0017)
TVAE	0.4185	0.7020	0.4989	0.8110	0.6829	0.9667
	(0.0061)	(0.0056)	(0.0061)	(0.0040)	(0.0268)	(0.0039)
CTGAN	0.4212	0.5090	0.2997	0.8193	0.7082	0.9638
	(0.0143)	(0.0127)	(0.0133)	(0.0050)	(0.0175)	(0.0022)
TabDDPM	0.5427	0.7771	0.6557	0.8476	0.7459	0.9862
	(0.0047)	(0.0025)	(0.0045)	(0.0014)	(0.0102)	(0.0012)
GReaT	0.4966	0.8042	-	-	0.7433	0.9697
	(0.0091)	(0.0024)			(0.0098)	(0.0070)

C STATISTICAL INFERENCE IN A CLASSIFICATION EXAMPLE

Preserving structure is crucial for statistical inference, not only in the regression case mentioned in Sec. 5, but also in classification. Here, we present a classification example where statistical inference plays a key role.

In this example, we estimate the decision boundary for object classification, a problem known as the support problem in the classification literature. To demonstrate that the EpBeta method results in a more robust boundary compared to existing mixup methods, we use a dataset with three classes distributed on a two-dimensional x-y plane. Specifically, we generate 500 instances for each class from the following distributions: $N\left([0,0]^\top, \mathbf{I}\right)$ for the first class, $N\left([2,0]^\top, \mathbf{I}\right)$ for the second class, and $N\left([4,0]^\top, \mathbf{I}\right)$ for the third class, where $\mathbf{I}$ is an identity matrix.

We then synthesize samples 100 times using mixup with EpBeta$(\delta = 0.05)$ and $\epsilon_0 = \epsilon_1 = 0.3$, or with Unif$(0,1)$. For each synthetic sample, we estimate the decision boundary using a support vector machine and calculate the intersection point of the boundary with the x-axis $(y = 0)$.

Table 9: Classification result.

	bias of boundary (class 1 vs 2)	bias of boundary (class 2 vs 3)	accuracy
EpBeta$(\delta = 0.05)$	+0.079 (0.062)	-0.048 (0.056)	0.798 (0.004)
Unif$(0,1)$	+0.249 (0.057)	-0.223 (0.049)	0.781 (0.004)

As shown numerically in Table 9, synthetic data generated by the EpBeta results in a more unbiased decision boundary compared to the uniform. This setting is actually related to the manifold intrusion problem (Guo et al., 2019), where the classification accuracy decreases for the second class, which is situated between the other two classes. This distortion affects not only statistical robustness, but also undermines classification accuracy.

D ADDITIONAL EXPERIMENT ON IMAGE DATA

Supervised contrastive learning (SupCL) (Khosla et al., 2020) is a powerful framework for learning effective representations for a variety of downstream tasks. However, it can lead to class-collapsed representations, where embedding outputs within the same class collapse to a single point, reducing performance (Islam et al., 2021; Chen et al., 2022; Lee et al., 2025). In other words, decreasing within-class variance of embedding outputs can harm performance.

To show the usefulness of EpBeta, we evaluate the transfer learning performances of SupCL. Specifically, we train the ResNet18 encoder with a 2-layer MLP projector head for 500 epochs, using a batch size of 500 and a temperature parameter of 0.1 in SupCL loss, on CIFAR-10 augmented with either EpBeta $(\delta = 0.05)$ or Uniform mixup. Following the transfer learning evaluation protocol (Kornblith et al., 2019; Lee et al., 2021), we remove the projector head and train a linear classifier on top of the frozen encoder using 6 downstream datasets: Dogs (Khosla et al., 2011), DTD (Cimpoi et al., 2014), Flowers (Nilsback and Zisserman, 2008), Food (Bossard et al., 2014), Pets (Parkhi et al., 2012), and MIT67 (Quattoni and Torralba, 2009). We repeat the entire process five times with different seeds. The table below presents the top-1 linear probing accuracy along with standard deviation. The results show that using EpBeta distribution instead of uniform distribution consistently outperforms, underscoring the benefits of variance preservation with EpBeta.

Table 10: Top-1 linear probing accuracy.

Name	Dogs	DTD	Flowers	Food	MIT67	Pets
EpBeta	0.150 (0.004)	0.404 (0.007)	0.552 (0.007)	0.301 (0.002)	0.366 (0.007)	0.268 (0.002)
Unif	0.140 (0.004)	0.390 (0.004)	0.540 (0.005)	0.284 (0.003)	0.355 (0.008)	0.246 (0.007)

E EPBETA PARAMETER EXAMPLES

Each cell in tables enumerates α and β in order that satisfy (17) in Theorem 10 and equality condition of (18) in Theorem 11 with $\alpha \geq \beta$ for given $\epsilon_0, \epsilon_1 \in \{0.0, 0.1, 0.2, \cdots, 0.9\}$, and $\delta \in \{0.005, 0.01\}$.

Table 11: Structure-preserving EpBeta parameters for $\delta = 0.005$.

	$\epsilon_1 = 0.0$	0.1	0.2	0.3	0.4
$\epsilon_0 = 0.0$	-	18.09, 1.91	33.16, 6.83	45.92, 14.08	56.86, 23.14
0.1	0.10, 0.01	20.17, 1.93	37.16, 6.96	51.72, 14.41	64.35, 23.80
0.2	0.20, 0.01	22.26, 1.96	41.18, 7.06	57.57, 14.69	71.90, 24.37
0.3	0.30, 0.01	24.35, 1.97	45.21, 7.16	63.44, 14.94	79.55, 24.88
0.4	0.41, 0.01	26.44, 1.99	49.24, 7.24	69.35, 15.16	87.23, 25.33
0.5	0.51, 0.01	28.54, 2.00	53.29, 7.31	75.28, 15.36	94.96, 25.73
0.6	0.61, 0.01	30.64, 2.02	57.36, 7.37	81.24, 15.54	102.74, 26.09
0.7	0.71, 0.01	32.73, 2.03	61.42, 7.43	87.22, 15.69	110.57, 26.42
0.8	0.81, 0.01	34.83, 2.04	65.49, 7.48	93.19, 15.84	118.39, 26.71
0.9	0.91, 0.01	36.94, 2.05	69.57, 7.53	99.23, 15.97	126.27, 26.99
	$\epsilon_1 = 0.5$	0.6	0.7	0.8	0.9
$\epsilon_0 = 0.0$	66.34, 33.67	74.62, 45.37	81.95, 58.06	88.44, 71.55	94.27, 85.75
0.1	75.38, 34.76	85.12, 47.03	93.79, 60.38	101.55, 74.66	108.50, 89.67
0.2	84.57, 35.74	95.82, 48.51	105.88, 62.46	114.95, 77.43	123.13, 93.25
0.3	93.86, 36.60	106.67, 49.83	118.17, 64.33	128.60, 79.94	138.06, 96.48
0.4	103.22, 37.37	117.61, 51.01	130.66, 66.03	142.49, 82.22	153.29, 99.45
0.5	112.69, 38.07	128.72, 52.09	143.27, 67.56	156.56, 84.30	168.81, 102.19
0.6	122.20, 38.69	139.88, 53.06	156.07, 68.98	170.87, 86.24	184.45, 104.66
0.7	131.81, 39.27	151.14, 53.95	168.97, 70.28	185.31, 88.01	200.32, 106.96
0.8	141.39, 39.78	162.52, 54.78	181.92, 71.45	199.86, 89.63	216.45, 109.13
0.9	151.09, 40.26	173.93, 55.53	194.99, 72.54	214.49, 91.11	232.58, 111.07

Table 12: Structure-preserving EpBeta parameters for $\delta = 0.01$.

	$\epsilon_1 = 0.0$	0.1	0.2	0.3	0.4
$\epsilon_0 = 0.0$	-	8.92, 0.99	16.49, 3.50	22.84, 7.15	28.28, 11.71
0.1	0.10, 0.04	9.98, 1.01	18.54, 3.57	25.79, 7.34	32.08, 12.07
0.2	0.21, 0.03	11.04, 1.02	20.59, 3.63	28.76, 7.49	35.91, 12.37
0.3	0.31, 0.03	12.11, 1.03	22.65, 3.69	31.76, 7.63	39.78, 12.64
0.4	0.41, 0.03	13.17, 1.04	24.72, 3.73	34.76, 7.75	43.68, 12.88
0.5	0.51, 0.03	14.23, 1.05	26.79, 3.78	37.79, 7.86	47.60, 13.10
0.6	0.62, 0.03	15.29, 1.06	28.86, 3.81	40.82, 7.96	51.55, 13.29
0.7	0.72, 0.03	16.35, 1.07	30.95, 3.85	43.86, 8.04	55.52, 13.47
0.8	0.82, 0.03	17.42, 1.07	33.03, 3.88	46.91, 8.12	59.49, 13.63
0.9	0.92, 0.03	18.48, 1.08	35.11, 3.90	49.96, 8.20	63.49, 13.77
	$\epsilon_1 = 0.5$	0.6	0.7	0.8	0.9
$\epsilon_0 = 0.0$	33.00, 17.00	37.12, 22.87	40.76, 29.23	44.00, 36.00	46.90, 43.11
0.1	37.57, 17.58	42.42, 23.74	46.73, 30.44	50.59, 37.60	54.06, 45.13
0.2	42.21, 18.09	47.82, 24.51	52.82, 31.52	57.33, 39.03	61.41, 46.96
0.3	46.91, 18.55	53.28, 25.20	59.03, 32.49	64.22, 40.33	68.95, 48.64
0.4	51.65, 18.95	58.83, 25.82	65.32, 33.37	71.23, 41.51	76.61, 50.15
0.5	56.44, 19.32	64.42, 26.37	71.69, 34.16	78.33, 42.58	84.41, 51.55
0.6	61.25, 19.65	70.08, 26.89	78.14, 34.89	85.53, 43.57	92.32, 52.84
0.7	66.11, 19.95	75.77, 27.35	84.64, 35.56	92.80, 44.48	100.31, 54.01
0.8	70.99, 20.23	81.51, 27.78	91.20, 36.17	100.13, 45.31	108.43, 55.12
0.9	75.88, 20.48	87.27, 28.17	97.81, 36.74	107.54, 46.09	116.59, 56.13