A Generalized Theory of Mixup for Structure-Preserving Synthetic Data

Chungpa Lee, Jongho Im, Joseph H.T. Kim

We investigate the statistical properties of synthetic data generated by Mixup. Our analysis reveals that **classical Mixup can distort key statistical characteristics**, such as (co)variance, leading to unintended consequences in downstream analysis. To address this issue, we propose **a generalized Mixup weighting scheme that can preserves the structure of the original data**. Numerical experiments confirm our theoretical findings and **mitigate concerns about model collapse**.

Synthetic Data Generated by Mixup

A continuous variable ($\tilde{X}$ or $\tilde{Y}$) is synthesized as a convex combination of the original values from two randomly selected instances (indexed by i and j). For a categorical variable ($\tilde{L}$), one of the two corresponding categories is randomly chosen:

Original instance: (X_i, Y_i, L_i) for $i \in [n]$
- $X_i, Y_i \in \mathbb{R}$: continuous variables
- $L_i \in [c]$: a categorical variable with c categories

Synthetic instance: $(\tilde{X}, \tilde{Y}, \tilde{L})$
- $\tilde{X} = W^X X_i + (1 - W^X) X_j$ and $\tilde{Y} = W^Y Y_i + (1 - W^Y) Y_j$
- $\tilde{L} = \begin{cases} L_i & \text{if } W^L \geq \tau \\ L_j & \text{if } W^L < \tau \end{cases}$

  where $\tau \in \mathbb{R}$ is the pre-defined cut point for the categorical variable, with $\tau = 0.5$ being a default choice.

Mixup Weight Scheme

Common mixup methods impose the same weight for each instance, i.e., $W^X = W^Y = W^L$.

Properties of Synthetic Data Generated by Mixup

Continuous Variable

The (co)variance of the synthesized variable is preserved, if and only if the first and second moments of the Mixup weight are equal.

Lemma 1 & Theorem 2. *For any synthetic pair* $(\tilde{X}, \tilde{Y})$,

1. $\mathrm{Cov}[\tilde{X}, \tilde{Y}] = \mathrm{Cov}[X, Y]$, *iff* $\mathbb{E}[W^X W^Y] = \frac{1}{2}(\mathbb{E}[W^X] + \mathbb{E}[W^Y])$.
2. $\mathrm{Var}[\tilde{X}] = \mathrm{Var}[X]$, *iff* $\mathbb{E}[(W^X)^2] = \mathbb{E}[W^X]$.
3. $\mathrm{Var}[\tilde{X}] < \mathrm{Var}[X]$, *iff* $\mathbb{E}[(W^X)^2] < \mathbb{E}[W^X]$.
4. $\mathrm{Var}[\tilde{X}] > \mathrm{Var}[X]$, *iff* $\mathbb{E}[(W^X)^2] > \mathbb{E}[W^X]$.

Corollary 3. *For any synthetic pair* $(\tilde{X}, \tilde{Y})$ *generated from* (X, Y) *using the weight constraint of* $W_k^X = W_k^Y$ *for all* $k \in [m]$, $\mathrm{Corr}[\tilde{X}, \tilde{Y}] = \mathrm{Corr}[X, Y]$.

Corollary 3 shows that the preservation of linear regression coefficients under equal instance-weight Mixup.

Continuous Variable Conditioned on Categorical Variable

Definition. *Given mixup weights* W^X *and* W^L *with a cut point* τ, *define* $u = \mathbb{E}\left[(1 - W^X)\mathbf{I}\{W^L \geq \tau\} + W^X \mathbf{I}\{W^L < \tau\}\right]$, *where* $\mathbf{I}$ *denotes an indicator function.*

Theorem 4. *For any synthetic pair* $(\tilde{X}, \tilde{L})$ *generated from* (X, L), *the synthetic conditional mean* $\mathbb{E}\left[\tilde{X} \middle| \tilde{L} = l\right]$ *can be expressed as*

$$\mathbb{E}\left[\tilde{X} \middle| \tilde{L} = l\right] = (1 - u) \cdot \mathbb{E}\left[X \middle| L = l\right] + u \cdot \mathbb{E}[X]$$
$$= (1 - u) \cdot \Pr\{L \neq l\} \cdot \mathbb{E}\left[X \middle| L = l\right] + u \cdot \Pr\{L \neq l\} \cdot \mathbb{E}[X \mid L \neq l].$$

Variance-Reduction and Structure-Preserving Mixup

Variance-Reduction Mixup

The mixup weight is commonly drawn from a distribution on $[0,1]$, such as the Beta or $\mathrm{Unif}(0,1)$ distribution. However this necessarily reduces variance, restoring the analysis.

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

Structure-Preserving Mixup

We propose the EpBeta distribution (*Definition 2* in our paper) as the mixup weight, which preserves (co)variance and controls conditional equivalence via a preset δ.

Theorem 10 & 11. *Given the small value* $\delta \geq 0$, *select parameters of the* EpBeta *distribution by Algorithm 1 in our paper. For an arbitrary synthetic triple* $(\tilde{X}, \tilde{Y}, \tilde{L})$ *generated from* (X, Y, L) *using the mixup weight of* $W^X = W^Y = W^L \sim \mathrm{EpBeta}$, *the followings are hold:*

1. $\mathrm{Var}[\tilde{X}] = \mathrm{Var}[X]$ *and* $\mathrm{Cov}[\tilde{X}, \tilde{Y}] = \mathrm{Cov}[X, Y]$
2. $\left|\mathbf{E}\left[\tilde{X} \middle| \tilde{L} = l\right] - \mathbf{E}\left[X \middle| L = l\right]\right| = O(\delta)$
3. $\left|\mathrm{Var}\left[\tilde{X} \middle| \tilde{L} = l\right] - \mathrm{Var}\left[X \middle| L = l\right]\right| \leq O(\delta)$

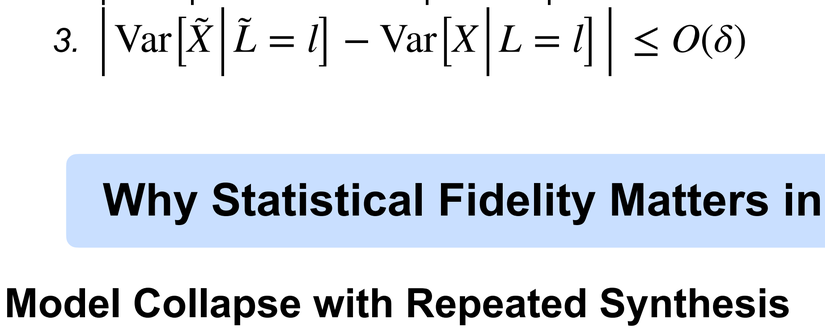
Why Statistical Fidelity Matters in Synthetic Data

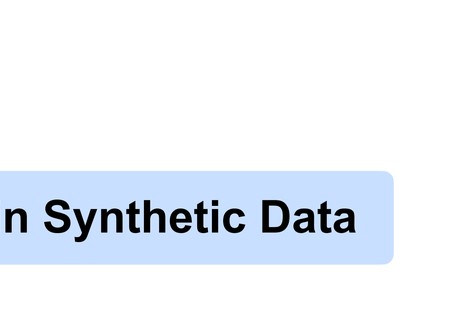
Model Collapse with Repeated Synthesis

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

Top-1 accuracy of CIFAR-10 classification models trained on repeatedly synthesized data, averaged over five runs using different synthetic datasets. After more than 20 iterations of resynthesis, **proposed** EpBeta **maintains stable performance by preserving the underlying data structure**, whereas Unif leads to significant drops in accuracy.

Relative Bias and Statistical Inference





Left panel: Relative bias of covariance (triangle) and expectation (bar) for the tabular dataset across synthetic generation methods, including GAN-based and diffusion-based models. Blue indicates negative bias, red positive, and grey unbiased. **Our method best preserves original statistical properties, enabling more reliable analysis.**

Right panel: In regression analysis **the proposed method consistently yields estimates close to the true value** β (red line), while the weight distribution sometimes leads to large deviations.