DIFFERENTIALLY PRIVATE VINE COPULAS FOR PRESERVING HIGHER-ORDER DEPENDENCE STRUCTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating high-utility synthetic data under differential privacy (DP) remains challenging when complex dependence and higher-order interactions are present. Traditional marginal and graph-based DP methods can preserve low-dimensional structure, such as univariate marginals and pairwise correlations, but often fail to recover higher-order interactions. While DP deep generative models such as GANs can, in principle, capture complex dependence, they are often unstable and difficult to tune. We propose Differentially Private Vine Copulas (DPRIVVE), an end-to-end differentially private synthetic data generator that models higher-order dependence via conditional propagation. DPRIVVE combines differentially private marginal transformations to construct DP pseudo-observations without data leakage, private D-vine ordering, and private copula family and parameter estimation. Notably, DPRIVVE more accurately captures higher-order interactions relative to competing DP baselines. Moreover, these improvements consistently hold over varying privacy budgets and dependence structures while preserving standard utility metrics.

1 INTRODUCTION

Differentially private (DP) synthetic data generation has become an important tool for sharing sensitive datasets while preserving formal privacy guarantees. Yet, existing DP generators are built almost exclusively for low-dimensional fidelity (e.g., univariate marginals and pairwise correlations) in spite of the analyses relying on these datasets often depending on more complex higher-order and conditional relationships among variables. A synthetic dataset matching the original across marginal structure exactly can still produce false scientific conclusions when the signal of interest lies in multi-way structures. Importantly, standard evaluation metrics fail to detect this, for instance, downstream classification accuracy and marginal-fidelity scores will remain stable even when interaction structure has been completely distorted.

The consequences of skewing conclusions can be quite impactful. For instance, large biobanks have been a massively successful resource for scientific discovery, but the high sensitivity of whole-genome sequencing makes data sharing difficult without formal privacy protection. Complex diseases, however, are often driven by higher-order interactions among genetic variants (Wang et al., 2025; Basu et al., 2018; Kumbier et al., 2018; Behr et al., 2022), and these low-signal effects can easily be wiped out via standard DP mechanisms. The same pattern arises in clinical research, where treatment effects are frequently mediated by combinations of comorbidities, or in social science, where outcomes depend on multi-way conditional relationships. In each case, a synthetic dataset that seems correct marginally can mislead the very analyses it was released to support.

To understand these challenges, note that methods built around low-order marginals or graphical-model factorizations implicitly compress multivariate dependence into local summaries. As interaction strength grows, this compression preserves marginal fidelity but attenuates the conditional and nonlinear relationships that govern multi-way structure. Classical approaches such as PrivBayes (Zhang et al., 2017), MST (McKenna et al., 2021), and AIM (McKenna et al., 2022) target low-order marginals or pairwise dependence as the optimization objective. Specifically, a mechanism with the objective to produce a set of marginals will, by construction, be indifferent to any two joint distributions that agree on those marginals. Meaning, the DP procedure has no reason to weight the true joint over an interaction-free one. Avoiding this hurdle requires a generative representation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where multi-way dependence is carried out explicitly rather than reconstructed post hoc from a set of low-order summaries.

Similarly, deep learning approaches, such as DP versions of generative adversarial networks (GANs) and other neural generators, should allow complex modeling in principle, but in practice are often expensive to train, sensitive to hyperparameters, and prone to instability (Saxena & Cao, 2021). Moreover, these methods may fail to preserve even simple dataset characteristics without careful tuning due to instability in the GAN training process, which can be problematic when an analysis requires accurate marginals and other low-dimensional statistics.

Existing methods are often only evaluated based on simplistic heuristics, such as classification error. While this is a common learning task, it ignores underlying structure misspecification, which in turn reduces the interpretability of the resulting models. In this work, we show directly how these methods fail in practice, even if utility-based metrics based on accuracy would fail to signal issues.

Motivated by the issues discussed, we develop DPRIVVE, a differentially private synthetic data generation algorithm based on vine copulas. Vine copulas provide a framework for multivariate modeling by decomposing a joint distribution into univariate marginals and a sequence of bivariate copulas arranged in a vine structure (Kim et al., 2013; Brechmann & Schepsmeier, 2013). This supplies the representation identified above: because each higher level of the vine is fit on pseudo-observations that have already been conditioned on the preceding variables, multi-way dependence is carried explicitly in the model rather than reconstructed from low-order summaries. Furthermore, it separates marginal behavior from dependence, so conditional structure can be improved without substantial harm to univariate fidelity. Our main contributions are: (i) we identify and characterize interaction collapse, a systematic failure of existing DP synthetic data generators to preserve higher-order structure despite strong low-order fidelity, and we show that standard evaluation metrics fail to detect this; (ii) we propose DPRIVVE, an interaction-preserving DP generation framework built on conditional dependence propagation via a privatized vine decomposition with an end-to-end pure ε-DP analysis and budget allocation; (iii) we demonstrate improved recovery of interaction structure across continuous and mixed-type simulated experiments and real tabular benchmarks, all while remaining competitive on standard fidelity, utility, and leakage metrics; and (iv) we provide practical usage guidance including default budget splits, and recommended sample-size/data-type regimes.

2 Background

2.1 Differential Privacy

Differential privacy (DP) provides a formal guarantee that the output of a randomized algorithm is insensitive to the inclusion or removal of any single individual's data (Dwork & Roth, 2014; Near & Abuah, 2021). Throughout this paper, datasets D and D' are considered neighboring if they differ in one record.

Definition 1 ((ε, δ)-Differential Privacy)**.** *A randomized mechanism M satisfies (ε, δ)-DP if for all neighboring datasets D, D' and measurable output sets S,*

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta.$$

Smaller values of ε correspond to stronger privacy guarantees, while δ allows a negligible probability of larger deviations.

We use standard ε-DP machinery throughout: the Laplace mechanism, which releases $f(D) + \mathrm{Lap}(\Delta f/\varepsilon)$ for a query of global ℓ_1 sensitivity $\Delta f = \max_{D \sim D'} \|f(D) - f(D')\|_1$; the exponential mechanism, which samples an output r with probability proportional to $\exp(\varepsilon\, u(D, r)/2\Delta u)$ for a utility u of sensitivity Δu; and the sequential-composition and postprocessing properties.

2.2 Vine Copulas

Copulas provide a method for multivariate modeling by separating marginal behavior from dependence.

Definition 2. ***Sklar's Theorem****. For a d-dimensional continuous random vector $X = (X_1, \ldots, X_d)$ with joint CDF F and marginal CDFs $F_1, \ldots, F_d$, Sklar's theorem (Sklar, 1959; Kim et al., 2013)*

states that there exists a copula $C : [0,1]^d \to [0,1]$ *such that*

$$F(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big).$$

The copula C captures the dependence structure among the components of X. Thus, multivariate modeling simplifies to modeling the marginals and the dependence separately.

Given observations $\{x_{ij}\}_{i=1}^n$, it is common to form pseudo-observations $u_{ij} = \hat{F}_j(x_{ij}) \in (0,1)$, which map each margin to approximately uniform variables. $z_{ij} = \Phi^{-1}(u_{ij})$.

While classical multivariate copulas (e.g., Gaussian, Student-t, and exchangeable Archimedean families) can be restrictive in higher dimensions, vine copulas (pair-copula constructions) provide an alternative by expressing a d-dimensional copula through a cascade of bivariate unconditional and conditional pair copulas.

In this paper, we focus on ***D-vines***, a widely used subclass of regular vines in which the associated graphical structure forms a path (Kim et al., 2013). A D-vine is specified by an ordering of variables $\pi = (\pi_1, \ldots, \pi_d)$, where dependence is modeled sequentially along the order and refined through higher-level conditional pair copulas.

Under the *simplifying assumption*, each conditional pair copula is assumed not to depend on the specific values of its conditioning variables, but only through the transformed conditional arguments (Brechmann & Schepsmeier, 2013). With this assumption, the D-vine copula density admits the pair-copula factorization

$$\begin{aligned} c(u_{\pi_1}, \ldots, u_{\pi_d}) = \prod_{k=1}^{d-1} \prod_{i=1}^{d-k} & c_{\pi_i, \pi_{i+k};\, \pi_{i+1:\pi_{i+k-1}}} \\ & \times \Big(u_{\pi_i|\pi_{i+1:\pi_{i+k-1}}},\, u_{\pi_{i+k}|\pi_{i+1:\pi_{i+k-1}}}\Big) \end{aligned} \tag{1}$$

where $c_{a,b;D}$ denotes the density of the pair copula between variables a and b conditional on the set D, and the conditional arguments are computed recursively.

Computation in vine models relies on copula h-*functions*, which correspond to conditional distribution functions implied by a bivariate copula. For a bivariate copula $C_{ab}(\cdot, \cdot; \theta)$ with parameter θ, the h-function is $h_{a|b}(u_a \mid u_b; \theta) = \frac{\partial}{\partial u_b} C_{ab}(u_a, u_b; \theta)$ and is used to obtain conditional pseudo-observations (e.g., $u_{a|D}$) that appear in higher tree levels.

2.3 Previous Work

The concept of integrating vine copulas within a DP framework was introduced by Gambs et al. (2021) in their method *COPULA-SHIRLEY*. However, later analysis by Nombo & Charest (2024) showed that the method does not provide full end-to-end differential privacy because downstream computations depend on both privatized quantities and the original data when constructing pseudo-observations. In contrast, DPRIVVE ensures that all downstream computations depend only on DP-released quantities. Other work has also explored DP copula-based modeling, including non-parametric copulas (Osorio-Marulanda et al., 2025) and scalable Gaussian copula approaches (Asghar et al., 2021).

3 Methods

3.1 DPRIVVE

Our proposed method, DPRIVVE, ensures DP by adding calibrated noise during multiple stages of the copula construction process to protect the privacy of the pseudo-observations. Below, we present an overview of the DPRIVVE algorithm.

3.1.1 DPRIVVE Overview

We describe each step of the DPRIVVE framework, explaining how DP is enforced throughout the pipeline and how the final synthetic data are generated by post-processing previously privatized quantities.

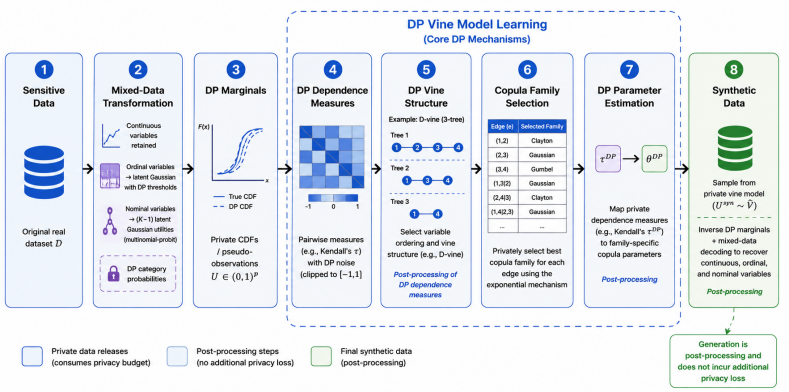


Figure 1: Overview of the DPRIVVE algorithm. The method privatizes multiple stages of vine copula construction to generate differentially private synthetic data.

1. **Sensitive data.** The algorithm begins with the original sensitive tabular dataset D.
2. **Mixed-data transformation.** To accommodate datasets containing continuous, ordinal, and nominal variables, we adopt a latent Gaussian representation for mixed-type data following Zhang et al. (2024). Continuous variables are retained on their original numerical scale. Ordinal variables are represented using a latent Gaussian variable, with thresholds determined from differentially private estimates of the corresponding category probabilities. For a nominal variable with K categories, we use a multinomial-probit representation with $K-1$ latent Gaussian utilities. The resulting continuous latent representation is then used for the subsequent marginal and vine-copula modeling steps.
3. **DP marginals.** For each variable j in the continuous latent representation, we construct a DP estimate of its marginal CDF, $\widehat{F}_j^{\mathrm{DP}}$, using a histogram mechanism. We discretize the variable into B bins over a prespecified range and compute the corresponding histogram counts. To ensure a fair comparison, we use the same discretization settings for continuous variables as in AIM. Since the ℓ_1-sensitivity of the histogram count vector is $\Delta = 1$ under add/remove adjacency, independent Laplace noise calibrated to the allocated marginal privacy budget is added to each bin count. Negative noisy counts are truncated to zero, and the resulting counts are normalized and cumulatively summed to obtain $\widehat{F}_j^{\mathrm{DP}}$. Each record is then mapped to the unit interval via

$$U_{ij} \leftarrow \widehat{F}_j^{\mathrm{DP}}(x_{ij}),$$

   yielding pseudo-observations $U \in (0,1)^p$ constructed from the DP marginal estimates.
4. **DP dependence measures.** We compute DP versions of Kendall's τ by adding Laplace noise with sensitivity $4/(n+1)$ to the empirical Kendall's τ estimates (Li et al., 2014). The noisy estimates are clipped to $[-1,1]$, which is permissible by the post-processing property of differential privacy.
5. **DP vine structure.** Using the privatized dependence measures, we construct the D-vine ordering. Specifically, we solve a traveling salesman problem on a complete graph with edge weights $d_{ij} = 1 - |\tau_{ij}^{\mathrm{DP}}|$. This encourages strongly dependent variables to appear close together in the D-vine ordering. Since the ordering is determined entirely from the privatized dependence measures, this step is post-processing and incurs no additional privacy loss.
6. **Copula family selection.** For each vine edge e, we privately select a pair-copula family using the exponential mechanism over a candidate family set $\mathcal{F}$. Given edge pseudo-observations $(u_i, v_i)_{i=1}^n$, each candidate family $f \in \mathcal{F}$ is scored using the clipped log-likelihood utility $u(f) = \sum_{i=1}^n \mathrm{clip}\left(\log c_f(u_i, v_i), -L, L\right)$. Since clipping bounds the sensitivity by $\Delta u = 2L$, the exponential mechanism samples family f with probability proportional to $\exp\left(\frac{\epsilon_{\mathrm{fam},e}\, u(f)}{2\Delta u}\right) = \exp\left(\frac{\epsilon_{\mathrm{fam},e}\, u(f)}{4L}\right)$, where $\epsilon_{\mathrm{fam},e}$ denotes the privacy budget allocated to family selection for edge e. The clipped log-likelihood values are used

only internally by the exponential mechanism and are not released; therefore, they do not require an additional privacy budget beyond $\epsilon_{\mathrm{fam},e}$.

7. **DP parameter estimation.** After selecting the copula family for each edge, we estimate the corresponding copula parameter using the privatized dependence information. Specifically, Kendall's τ^{DP} is mapped to the corresponding copula parameter using the family-specific relationship between Kendall's τ and the copula parameter. For example, for a Gaussian copula,
$$\rho = \sin\Big(\frac{\pi}{2}\tau^{\mathrm{DP}}\Big),$$
while analogous mappings are used for the other candidate copula families. Since parameter estimation relies only on previously privatized dependence measures, it is post-processing and incurs no additional privacy loss.
8. **Synthetic data generation.** Finally, we sample synthetic pseudo-observations from the fitted private vine model, $U^{\mathrm{syn}} \sim \widehat{\mathcal{V}}$, and map them back through the DP marginal inverses: $X_j^{\mathrm{syn}} \leftarrow \left(\widehat{F}_j^{\mathrm{DP}}\right)^{-1}\left(U_j^{\mathrm{syn}}\right)$. For mixed-type variables, the resulting synthetic latent variables are subsequently decoded to recover variables on their original continuous, ordinal, or nominal scales. Since synthetic data generation and decoding depend only on previously privatized quantities, these operations are post-processing and incur no additional privacy loss.

3.2 Privacy Guarantee and Accounting

Setup and neighboring datasets. Recall our previous definition of neighboring datasets: datasets D and D' are neighbors if they differ in one row. Let n denote the number of records, p the number of variables, and $t_{\max}$ the truncation level of the D-vine.

Budget split. Fix a total privacy budget $\epsilon > 0$ and allocate it across stages as
$$\epsilon = \epsilon_{\mathrm{cat}} + \epsilon_{\mathrm{marg}} + \epsilon_{\mathrm{ord}} + \epsilon_{\tau} + \epsilon_{\mathrm{fam}},$$
where: (i) ϵ_{cat} is used to estimate category probabilities for the mixed-data transformation, (ii) ϵ_{marg} is used for DP marginals, (iii) ϵ_{ord} is used for the DP dependence measures used to determine the D-vine ordering, (iv) ϵ_{τ} is used for DP Kendall's τ estimates on vine edges, and (v) ϵ_{fam} is used for exponential-mechanism family selection on vine edges.

Theorem 1 (Privacy of DPRIVVE)**.** *Under the neighboring dataset definition, the* DPRIVVE *mechanism is ϵ-differentially private with*
$$\epsilon = \epsilon_{\mathrm{cat}} + \epsilon_{\mathrm{marg}} + \epsilon_{\mathrm{ord}} + \epsilon_{\tau} + \epsilon_{\mathrm{fam}},$$
where the five terms correspond to the mixed-data transformation, DP marginals, DP ordering statistics, DP vine-edge statistics, and DP family selection, respectively.

3.3 Scalability

We acknowledge that the computational cost of vine copulas can grow substantially with the number of variables and observations. Moreover, in the differentially private setting, estimating a large number of pairwise dependence quantities requires the available privacy budget to be distributed across many components. For higher-dimensional settings, DPRIVVE can be scaled by partitioning variables into smaller dependence blocks and fitting separate vine copulas within each block, thereby reducing the number of pairwise quantities and pair-copula components that must be estimated. When the block structure is prespecified or obtained from public information, this partitioning incurs no additional privacy cost; data-adaptive blocking based on the private data would instead require a separately accounted privacy mechanism. DPRIVVE can also be truncated at an earlier vine-tree level, directly reducing the number of conditional pair-copula components that must be privately estimated. Both strategies reduce computational requirements and allow the available privacy budget to be distributed across fewer dependence components, although blockwise modeling may omit dependence across blocks and aggressive truncation may discard relevant conditional dependence.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We evaluate the DP algorithms on four real-world datasets and two controlled higher-order simulation settings, summarized in Table 1.

Table 1: Summary of datasets used in the experiments.

Data set	Obs.	Numeric	Binary	Ordinal	Nominal
Wine	6,497	11	1	0	0
California Housing	20,640	9	0	0	0
Pima Indian (Diabetes)	768	8	1	0	0
Framingham	4,238	8	7	1	0
Simulated (continuous)	20,000	4	0	0	0
Simulated (mixed-type)	20,000	3	0	0	1

4.1.1 REAL-WORLD DATA

We evaluate our methods on four real-world datasets: Wine Quality, California Housing, Pima Indian Diabetes, and the Framingham Heart Study. Wine Quality, California Housing, and Pima Indian Diabetes provide primarily numerical benchmark settings, while Framingham provides a mixed-type setting containing continuous, binary, and ordinal variables. For the California Housing dataset, we convert median house value into a binary outcome indicating whether a house value is above or below the median.

4.1.2 SIMULATED DATA

Continuous higher-order dependence simulation. To evaluate each DP method's ability to preserve higher-order dependence, we consider a controlled simulation in which three continuous predictors (x_1, x_2, x_3) are drawn from a multivariate normal distribution with mean zero and unit marginal variances. We consider two dependence structures. In the *constant correlation* setting, all pairwise correlations are $\rho = 0.7$. In the *non-constant correlation* setting,

$$(\rho_{12}, \rho_{13}, \rho_{23}) = (0.70,\ 0.50,\ 0.30).$$

Given (x_1, x_2, x_3), we generate a fourth continuous variable according to

$$x_4 = 0.6x_1 + 0.6x_2 + 0.6x_3 + \beta_{\text{int}}(x_1x_2x_3) + \varepsilon, \qquad \varepsilon \sim N(0, 0.5^2).$$

The main-effect terms induce linear dependence between x_4 and the individual predictors, while the multiplicative term introduces a three-way interaction that is not captured by pairwise dependence alone. We use $n = 20{,}000$ observations, vary the interaction strength over

$$\beta_{\text{int}} \in \{0, 0.5, 1, 2, 3, 5\},$$

and evaluate each method across privacy budgets $\epsilon \in \{1, 2, 3, 4, 5\}$.

Mixed-type higher-order simulation. We additionally consider $n = 20{,}000$ observations with continuous variables (x_1, x_2, y) and a three-level nominal variable $G \in \{\texttt{Low}, \texttt{Mid}, \texttt{High}\}$. We generate $\text{Cor}(x_1, x_2) = 0.6$ and construct G from a latent score depending on both predictors. The outcome includes the three-way interaction $\beta_{\text{int}}x_1x_2\mathbb{I}(G = \texttt{High})$, with $\beta_{\text{int}} \in \{0, 1, 3, 5\}$ and $\epsilon \in \{1, 2, 3, 4, 5\}$. Full details are provided in Appendix A.1.

4.2 DP ALGORITHMS

We compare DPRIVVE against several differentially private synthetic data baselines: MST (McKenna et al., 2021), PrivBayes (Zhang et al., 2017), AIM (McKenna et al., 2022), and DP-WGAN (Xie et al., 2018). MST, PrivBayes, and AIM represent widely used marginal and

graphical-model-based approaches for DP synthetic data generation, while DP-WGAN represents a deep generative model based on adversarial neural networks. Recent benchmarking work has found that marginal-based methods consistently outperform GAN-based approaches for tabular differentially private synthetic data generation, while GAN-based methods often struggle to preserve even low-dimensional statistics accurately (Tao et al., 2021). We include DP-WGAN in our comparisons to provide coverage of deep generative adversarial approaches and ensure that our evaluation spans both classical statistical and neural synthetic data generation methods.

All methods we compare to DPRIVVE are configured with comparable privacy budgets (ϵ, δ) and trained under identical data partitions to ensure consistency in privacy–utility evaluation. We set $\delta = 10^{-5}$ for all experiments. For DP-WGAN, we used a noise multiplier of $\sigma = 2.0$, sample rate 0.01, five critic updates per generator step, learning rate 10^{-4}, and weight clipping of 0.05. The number of training epochs was adjusted to satisfy the target privacy budgets. For AIM, we used a maximum marginal degree of 3.

4.3 Utility and Empirical Data Leakage Evaluation

4.3.1 Higher-Order Dependence

Simulated datasets. We evaluate higher-order fidelity by fitting $x_4 \sim x_1 * x_2 * x_3$ to the original and synthetic data and measuring recovery of the three-way interaction. We report $|\hat{\beta}_{123}^{\text{real}} - \hat{\beta}_{123}^{\text{synth}}|$ and the difference in holdout RMSE between models trained on synthetic and real data. For the mixed-type simulation, we analogously evaluate recovery of the three-way interaction in $y \sim x_1 * x_2 * G$.

Real datasets. Because the true higher-order structure is unknown in real data, we use decision-tree splitting behavior as a proxy for dependence preservation (Basu et al., 2018; Wang et al., 2025; Kumbier et al., 2018; Agarwal et al., 2023; Huang et al., 2025). We fit 100 bootstrap trees and compare the proportion of trees in which each feature is selected for a split between the original and synthetic data. We focus on the Framingham, Wine, Diabetes studies due to their stable tree structures, while the Housing dataset is used primarily to verify preservation of lower-order dependence and standard fidelity metrics. Additional implementation details are provided in the Appendix.

4.3.2 Standard Utility and Privacy Metrics

To verify that improved higher-order fidelity does not come at the expense of standard synthetic-data utility, we additionally evaluate marginal fidelity using KS and TVD, pairwise dependence using Pearson's r, Cramer's V, correlation ratio η, and CMD (Lautrup et al., 2024), and downstream predictive utility using logistic regression and random forests trained on synthetic data and evaluated on held-out real data. We additionally assess empirical data leakage using distance to closest record (DCR) and ε-identifiability risk. Full metric definitions and additional results are provided in the Appendix.

5 Results

5.1 Higher-Order Dependence

Simulated datasets. We present our simulation results for benchmarking how accurately the competing DP algorithms capture higher-order interactions. Below, we present results for the continuous data non-constant correlation setup.

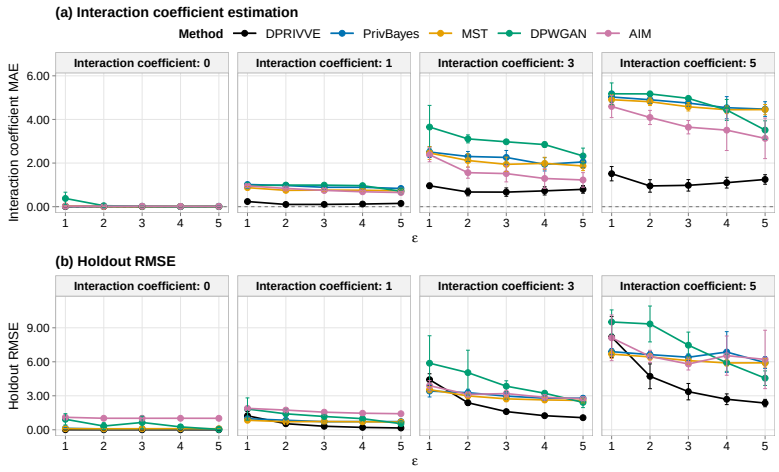


Figure 2: Higher-order dependence preservation under the non-constant correlation setting. Panel (a) shows the absolute error in the estimated three-way interaction coefficient, and Panel (b) shows holdout RMSE for models trained on synthetic data and evaluated on the real holdout set.

Across both correlation settings, all methods perform similarly when the interaction strength is small (e.g., $\beta_{\text{int}} \in \{0, 1\}$). However, as β_{int} increases from 0 to 5, MST, PrivBayes, DPWGAN, and AIM degrade across metrics, indicating difficulty capturing stronger higher-order structure. In contrast, DPRIVVE remains comparatively robust, maintaining lower interaction coefficient error and better holdout RMSE as interaction strength increases. We also observed substantial instability in GAN training, consistent with the larger variance and occasional failures of DPWGAN. Averaged over ϵ and interaction strengths, DPRIVVE reduces the mean absolute error of the recovered interaction coefficients by 78.42–82.24% in the constant-correlation setting and by 72.08–75.12% in the nonconstant-correlation setting relative to the other baselines. Similar results are shown in the mixed-data simulation setting. The figures for the constant and mixed-data correlation settings can be found in A.3

Real datasets. In the WINE study, DPRIVVE preserves the prominent role of alcohol and volatile.acidity observed in the original data, while several baselines exhibit larger changes in the feature-splitting profile (Figure 3). In particular, DPRIVVE retains high split frequencies for both variables, whereas PRIVBAYES substantially underrepresents them and DPWGAN produces very low split frequencies for the displayed features. These results suggest that DPRIVVE preserves aspects of the multivariate structure relevant to tree-based prediction that are not fully characterized by marginal and pairwise summaries. Additional results for the DIABETES and FRAMINGHAM datasets, as well as results across privacy budgets, are provided in the Appendix. On the mixed-type FRAMINGHAM dataset, DPRIVVE does not achieve the lowest tree-proxy error, but remains among the stronger-performing methods despite the additional challenge of modeling mixed continuous and categorical variables.

5.2 LOW-ORDER FIDELITY, PREDICTIVE UTILITY, AND PRIVACY LEAKAGE

Across standard synthetic-data evaluation metrics, DPRIVVE remains competitive with existing DP baselines. In particular, DPRIVVE achieves comparable predictive utility and univariate fidelity while generally providing strong preservation of pairwise dependence. In the mixed-type Framingham setting, DPRIVVE achieves particularly strong marginal fidelity while maintaining useful downstream predictive performance, although its pairwise association error is higher than several competing methods. We additionally find no evidence that its improved preservation of higher-order structure comes at the cost of increased empirical privacy leakage, as measured by ϵ-identifiability risk and distance to closest record (DCR). Thus, while DPRIVVE is not uniformly optimal on every conventional metric, its gains in higher-order dependence preservation are achieved while maintaining competitive standard utility and empirical privacy performance. Full results are provided in Appendix A.6.

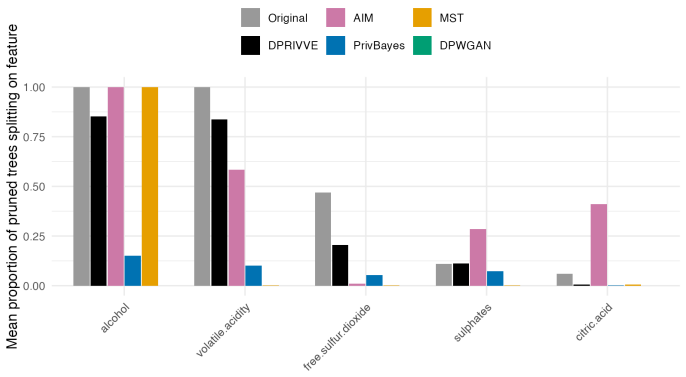


Figure 3: Decision-tree feature-splitting behavior for the WINE dataset at $\epsilon = 5$. Bars show the mean proportion of pruned bootstrap trees in which each of the five most frequently selected features from the original data appears as a splitting variable.

6 CONCLUSION

This work introduced DPRIVVE, a differentially private synthetic data generation method based on vine copulas, designed to preserve both low-order structure (e.g., univariate marginals and pairwise dependence) and higher-order interactions. Across the real-data benchmarks, DPRIVVE achieves utility that is competitive with widely used classical baselines such as MST and PrivBayes, demonstrating that a copula-based approach can match the strengths of graphical-model methods while also capturing multivariate dependence. In contrast, deep generative approaches such as GAN-based methods, while powerful in principle, were less reliable in our setting.

Beyond strong results on standard fidelity and downstream utility metrics, the primary advantage of DPRIVVE is its ability to better capture higher-order dependence. The higher-order interaction experiments highlight that DPRIVVE can more faithfully reproduce multiway dependence patterns than methods driven primarily by low-order marginals. This means DPRIVVE is strong for applications where analysis depends on interactions among multiple variables.

Capturing higher-order interactions and dependence is vital in real-world applications such as medicine and public health. Clinical outcomes are often influenced by complex combinations of different covariates, and these relationships are frequently nonlinear and conditional. Synthetic datasets that preserve only low-order dependence may appear realistic based on low-level summaries while failing to capture meaningful multi-way dependencies, potentially undermining important analysis.

6.1 LIMITATIONS

Several limitations should be noted. First, computational and privacy costs grow with dimensionality due to the large number of pairwise dependence estimates required for vine construction. Second, copula-based approaches are generally more reliable for continuous or predominantly continuous data and may degrade in heavily categorical settings, where copulas are not uniquely identified under discrete marginals. Third, DPRIVVE adopts the simplifying assumption commonly used in vine-copula modeling, under which conditional pair-copulas are assumed not to vary with the values of their conditioning variables. This assumption improves computational tractability but may limit flexibility when conditional dependence varies substantially across the conditioning space. Fourth, higher-tree conditional dependence estimation can become unstable in very small-sample regimes. Finally, performance may depend on design choices such as marginal discretization, vine truncation levels, and privacy-budget allocation. We examine sensitivity to these choices under the continuous nonconstant-correlation simulation setting in Appendix A.7.

REFERENCES

Abhineet Agarwal, Ana M Kenney, Yan Shuo Tan, Tiffany M Tang, and Bin Yu. Mdi+: A flexible random forest-based feature importance framework. *arXiv preprint arXiv:2307.01932*, 2023.

Hassan Jameel Asghar, Ming Ding, Thierry Rakotoarivelo, Sirine Mrabet, and Dali Kaafar. Differentially private release of datasets using gaussian copula. *Journal of Privacy and Confidentiality*, 10(2), Feb. 2021. doi: 10.29012/jpc.686. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/686.

Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences*, 119(22): e2118636119, 2022.

Eike Christian Brechmann and Ulf Schepsmeier. Modeling dependence with c- and d-vine copulas: The r package cdvine. *Journal of Statistical Software*, 52(3):1–27, 2013. doi: 10.18637/jss.v052.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v052i03.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Sébastien Gambs, Frédéric Ladouceur, Antoine Laurent, and Alexandre Roy-Gaumond. Growing synthetic data through differentially-private vine copulas. *Proceedings on Privacy Enhancing Technologies*, 2021(3):122–141, 2021. doi: 10.2478/popets-2021-0040.

Melody Huang, Tiffany M Tang, and Ana M Kenney. Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference. *arXiv preprint arXiv:2502.07275*, 2025.

Daeyoung Kim, Jong-Min Kim, Shu-Min Liao, and Yoon-Sung Jung. Mixture of d-vine copulas for modeling dependence. *Computational Statistics Data Analysis*, 64:1–19, 2013. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.02.018. URL https://www.sciencedirect.com/science/article/pii/S0167947313000741.

Karl Kumbier, Sumanta Basu, Erwin Frise, Susan E Celniker, James B Brown, Susan Celniker, and Bin Yu. Signed iterative random forests to identify enhancer-associated transcription factor binding. *arXiv preprint arXiv:1810.07287*, 2018.

Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1), December 2024. ISSN 1573-756X. doi: 10.1007/s10618-024-01081-4. URL http://dx.doi.org/10.1007/s10618-024-01081-4.

Haoyue Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pp. 475–486. OpenProceedings.org, 2014. doi: 10.5441/002/edbt.2014.43.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *CoRR*, abs/2108.04978, 2021. URL https://arxiv.org/abs/2108.04978.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11):2599–2612, July 2022. ISSN 2150-8097. doi: 10.14778/3551793.3551817. URL https://doi.org/10.14778/3551793.3551817.

Joseph P. Near and Chiké Abuah. *Programming Differential Privacy*, volume 1. 2021. URL https://programming-dp.com/.

Leila Nombo and Anne-Sophie Charest. Inference with combining rules from multiple differentially private synthetic datasets, 2024. URL https://arxiv.org/abs/2405.04769.

Pablo A. Osorio-Marulanda, John Esteban Castro Ramirez, Mikel Hernández Jiménez, Nicolas Moreno Reyes, and Gorka Epelde Unanue. Differentially private non parametric copulas: Generating synthetic data with non parametric copulas under privacy guarantees. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 441–448, 2025. doi: 10.1109/CBMS65348.2025.00094.

Divya Saxena and Jiannong Cao. Generative adversarial networks (gans): Challenges, solutions, and future directions. *ACM Comput. Surv.*, 54(3), May 2021. ISSN 0360-0300. doi: 10.1145/3446374. URL https://doi.org/10.1145/3446374.

Martin J. Sklar. Fonctions de repartition a n dimensions et leurs marges. 1959. URL https://api.semanticscholar.org/CorpusID:127105744.

Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. *CoRR*, abs/2112.09238, 2021. URL https://arxiv.org/abs/2112.09238.

Q. Wang, T. M. Tang, M. Youlton, et al. Epistasis regulates genetic control of cardiac hypertrophy. *Nat Cardiovasc*, 2025.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *CoRR*, abs/1802.06739, 2018. URL http://arxiv.org/abs/1802.06739.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017. ISSN 0362-5915. doi: 10.1145/3134428. URL https://doi.org/10.1145/3134428.

Yanqing Zhang, Qi Xu, Niansheng Tang, and Annie Qu. Differentially private data release for mixed-type data via latent factor models. *Journal of Machine Learning Research*, 25(116):1–37, 2024. URL http://jmlr.org/papers/v25/22-1324.html.

A Appendix

A.1 Mixed-Type Higher-Order Simulation Details

To evaluate DPRIVVE's ability to preserve higher-order dependence in a mixed-type setting, we generate a simulated dataset with $n = 20{,}000$ observations containing three continuous variables, x_1, x_2, and y, and one three-level nominal variable, G.

We first generate the two continuous predictors from a bivariate normal distribution,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \right).$$

We then generate a latent score for the categorical variable according to

$$Z_G = 0.8x_1 + 0.8x_2 + \varepsilon_G, \qquad \varepsilon_G \sim N(0, 1).$$

The latent score is divided at its empirical tertiles to construct the nominal variable

$$G \in \{\texttt{Low}, \texttt{Mid}, \texttt{High}\}.$$

This construction induces dependence between the categorical variable and both continuous predictors.

The continuous outcome is generated according to

$$y = 0.6x_1 + 0.6x_2 + 0.5\,\mathbb{I}(G = \texttt{Mid}) + 1.0\,\mathbb{I}(G = \texttt{High}) + \beta_{\text{int}} x_1 x_2 \mathbb{I}(G = \texttt{High}) + \varepsilon_y,$$

where

$$\varepsilon_y \sim N(0,1).$$

The term

$$x_1 x_2 \mathbb{I}(G = \texttt{High})$$

defines a three-way interaction involving two continuous variables and one nominal variable. We vary the interaction strength over

$$\beta_{\text{int}} \in \{0, 1, 3, 5\}.$$

For each simulated dataset, we fit

$$y \sim x_1 * x_2 * G$$

to both the original and synthetic training data. We evaluate higher-order fidelity using the absolute error in the estimated three-way interaction coefficient corresponding to the `High` category,

$$\left| \hat{\beta}_{\text{int}}^{\text{real}} - \hat{\beta}_{\text{int}}^{\text{synth}} \right|,$$

as well as sign recovery and predictive performance on the original holdout data using RMSE and R^2.

This simulation therefore provides a controlled benchmark for evaluating whether DPRIVVE preserves higher-order dependence that spans both continuous and categorical variable types.

A.2 PROOFS

Proof. Let D and D' be neighboring datasets. We track privacy loss across the randomized components of DPRIVVE; all remaining computations are post-processing of previously privatized quantities.

Step 1 (DP mixed-data transformation). For each categorical or ordinal variable j, category probabilities are estimated from Laplace-noised category counts. The category-count vector has ℓ_1-sensitivity 1, so releasing the privatized category probabilities with privacy budget $\epsilon_{\text{cat},j}$ is $\epsilon_{\text{cat},j}$-differentially private. By sequential composition across the categorical and ordinal variables,

$$\sum_j \epsilon_{\text{cat},j} = \epsilon_{\text{cat}}.$$

Construction of the corresponding latent Gaussian representation from these privatized category probabilities is post-processing and incurs no additional privacy loss.

Step 2 (DP marginals). For each variable j, the marginal estimate is obtained from a Laplace-noised histogram with ℓ_1-sensitivity 1 and privacy budget $\epsilon_{\text{marg},j}$. Hence, by the Laplace mechanism, the release for variable j is $\epsilon_{\text{marg},j}$-DP. By sequential composition across the variables,

$$\sum_{j=1}^{p} \epsilon_{\text{marg},j} = \epsilon_{\text{marg}}.$$

Constructing pseudo-observations from these DP marginal estimates is post-processing and incurs no additional privacy loss.

Step 3 (DP ordering). For each pair (i, j) used in the ordering stage, Kendall's τ has sensitivity

$$\Delta_\tau = \frac{4}{n+1},$$

as shown in (Li et al., 2014). Releasing a noisy dependence estimate with privacy budget $\epsilon_{\text{ord},ij}$ is therefore $\epsilon_{\text{ord},ij}$-DP by the Laplace mechanism. By sequential composition over all pairwise dependence measures used in this stage,

$$\sum_{i<j} \epsilon_{\text{ord},ij} = \epsilon_{\text{ord}}.$$

Computing the D-vine ordering from these privatized dependence measures is deterministic post-processing and incurs no additional privacy loss.

Step 4 (DP vine edges). For each vine edge e, we: (i) compute a DP Kendall's τ estimate with privacy budget $\epsilon_{\tau,e}$ and (ii) select a copula family using the exponential mechanism with privacy budget $\epsilon_{\mathrm{fam},e}$.

The DP Kendall's τ estimate is $\epsilon_{\tau,e}$-DP by the Laplace mechanism. For family selection, the clipped log-likelihood utility has sensitivity bounded by $\Delta u = 2L$, so the exponential mechanism is $\epsilon_{\mathrm{fam},e}$-DP for edge e.

By sequential composition across all m vine edges and both randomized mechanisms,

$$\sum_{e=1}^{m} \epsilon_{\tau,e} + \sum_{e=1}^{m} \epsilon_{\mathrm{fam},e} = \epsilon_{\tau} + \epsilon_{\mathrm{fam}}.$$

Step 5 (Parameter estimation, propagation, and synthetic data generation). Copula parameter estimation from privatized Kendall's τ, conditional propagation through h-functions, sampling from the fitted vine, inverse transformation through the DP marginal estimates, and decoding of the mixed-type latent variables are all functions of previously privatized quantities. Therefore, by the post-processing property of differential privacy, these operations incur no additional privacy loss.

Combining the privacy losses from Steps 1–4 by sequential composition gives

$$\epsilon_{\mathrm{cat}} + \epsilon_{\mathrm{marg}} + \epsilon_{\mathrm{ord}} + \epsilon_{\tau} + \epsilon_{\mathrm{fam}} = \epsilon.$$

Thus, DPRIVVE satisfies ϵ-differential privacy. □

A.3 HIGHER-ORDER SIMULATION RESULTS

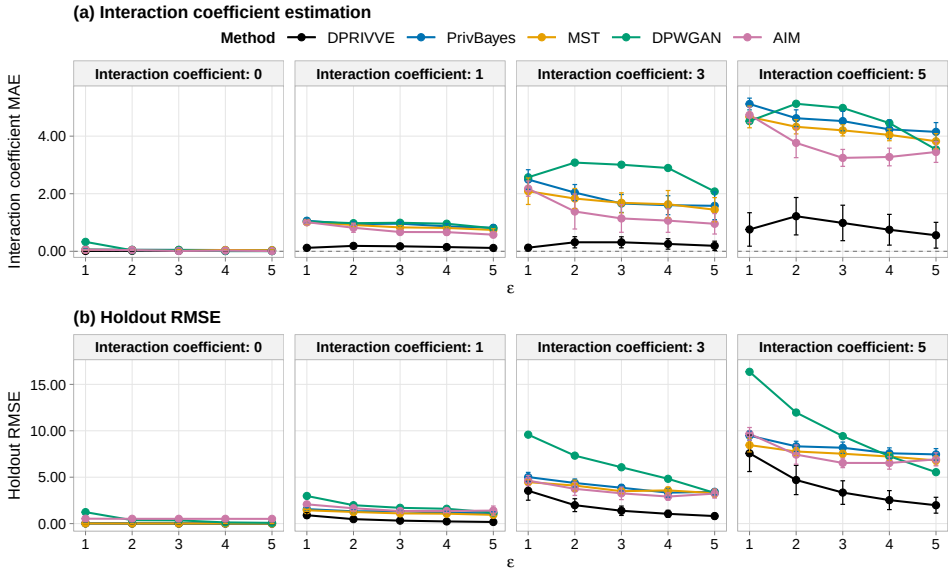


Figure 4: Higher-order dependence preservation in the constant-correlation simulation setting. Panel (a) shows the mean absolute error in the estimated three-way interaction coefficient across privacy budgets. Panel (b) shows the holdout RMSE for models trained on synthetic data and evaluated on the real holdout set. Error bars represent variability across simulation replicates.

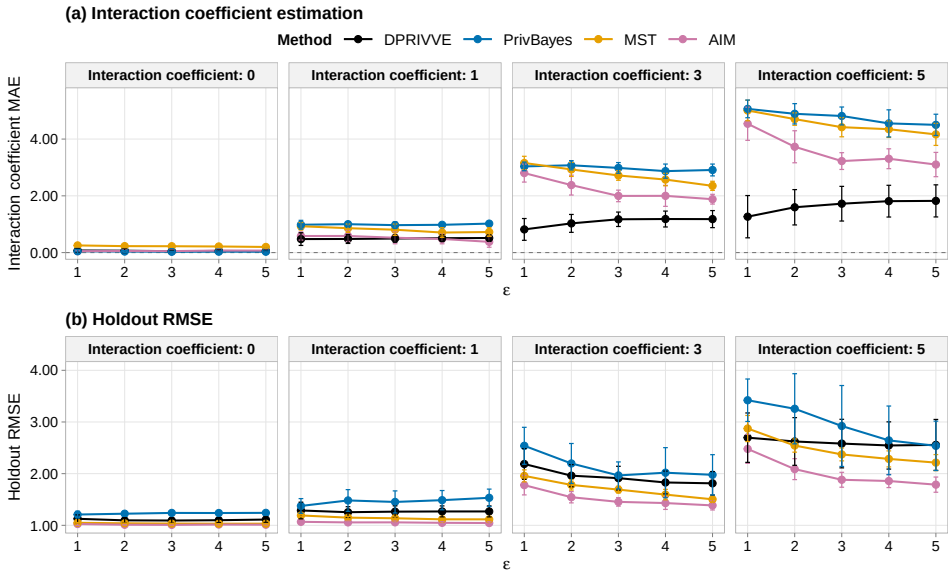


Figure 5: Higher-order dependence preservation in the mixed-data simulation setting. Panel (a) shows the mean absolute error in the estimated three-way interaction coefficient across privacy budgets. Panel (b) shows the holdout RMSE for models trained on synthetic data and evaluated on the real holdout set. Error bars represent variability across simulation replicates.

A.4 Additional Tree-Based Dependence Results

Figure 6 presents additional tree-based dependence results for the Diabetes dataset. DPRIVVE reproduces several features of the original splitting profile, including the prominent roles of glucose and BMI.

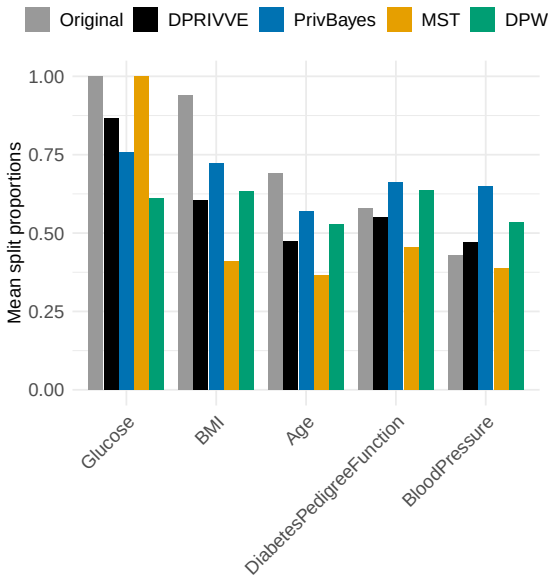


Figure 6: Decision-tree feature-splitting behavior for the Diabetes dataset at $\epsilon = 5$.

Figure 7 summarizes the corresponding tree-based results for the Framingham dataset at $\epsilon = 5$. DPRIVVE achieved a split-frequency MAE of 0.105, compared with 0.127 for PrivBayes, 0.124 for MST, 0.059 for AIM, and 0.150 for DPWGAN.

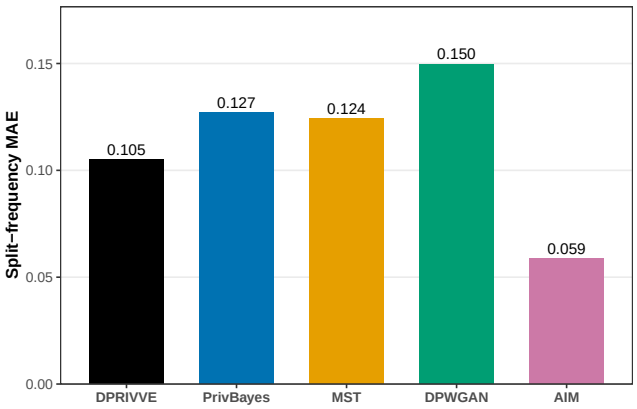


Figure 7: Tree-based dependence preservation for the Framingham dataset at $\epsilon = 5$. Bars show the mean absolute error between the synthetic and real-data feature split-frequency profiles, where lower values indicate closer agreement. Each method is evaluated relative to the real training sample used to fit the corresponding synthesizer.

A.5 Additional Machine Learning Results

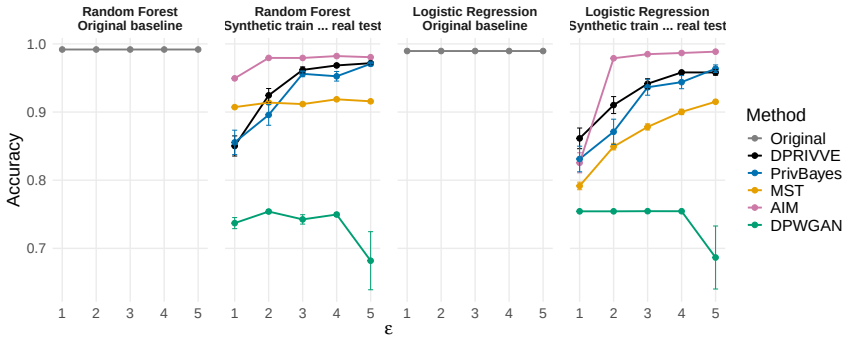


Figure 8: Classification accuracy for the WINE dataset. The original real-data baseline is shown in gray, and synthetic-data methods are evaluated by training on synthetic data and testing on the real holdout set.

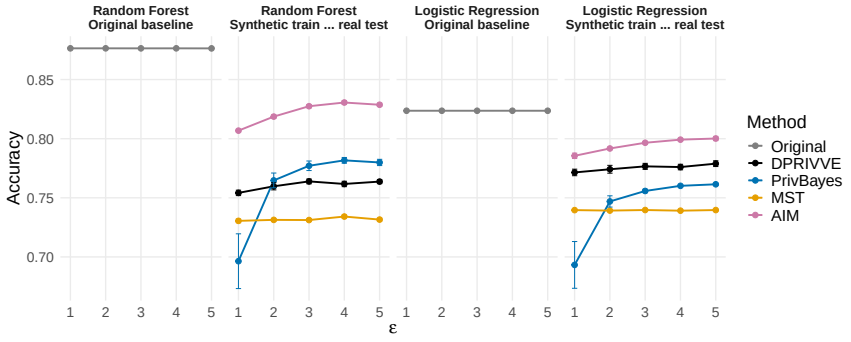


Figure 9: Classification accuracy for the HOUSING dataset. The original real-data baseline is shown in gray, and synthetic-data methods are evaluated by training on synthetic data and testing on the real holdout set.

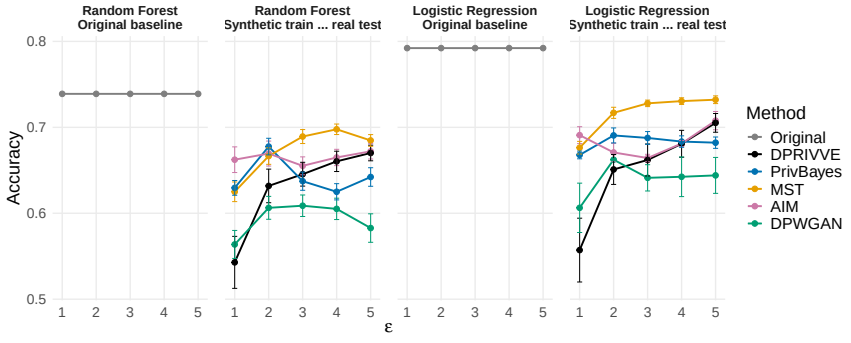


Figure 10: Classification accuracy for the DIABETES dataset. The original real-data baseline is shown in gray, and synthetic-data methods are evaluated by training on synthetic data and testing on the real holdout set.

A.6 Additional Fidelity Results

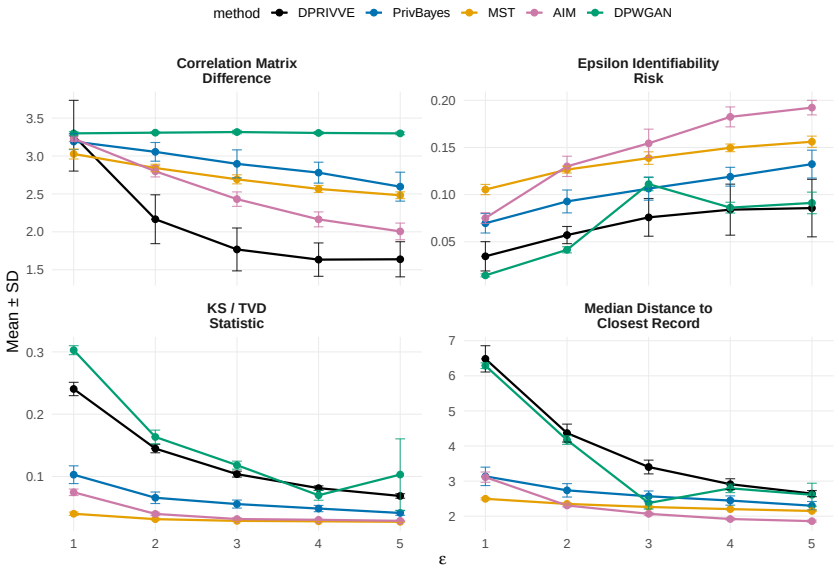


Figure 11: Additional fidelity and empirical privacy results for the WINE dataset across privacy budgets. Points show means and error bars show one standard deviation across synthetic-data replicates.

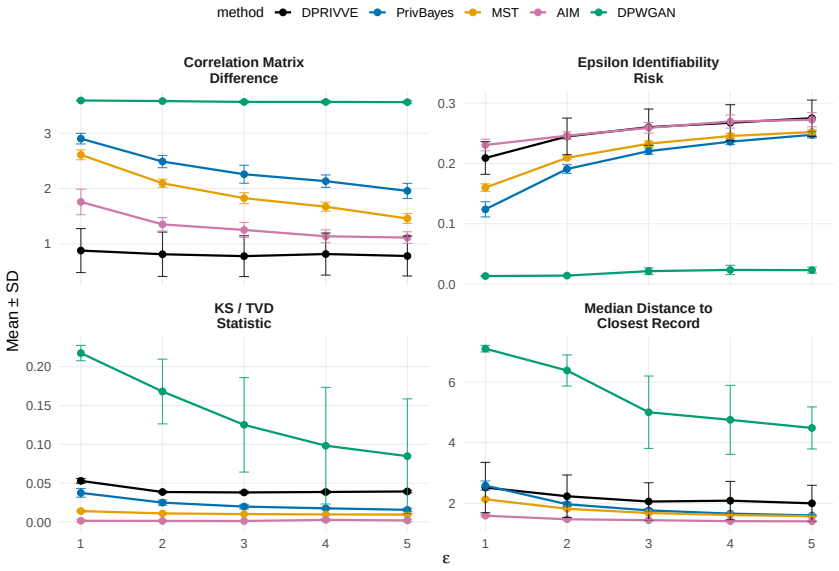


Figure 12: Additional fidelity and empirical privacy results for the HOUSING dataset across privacy budgets. Points show means and error bars show one standard deviation across synthetic-data replicates.

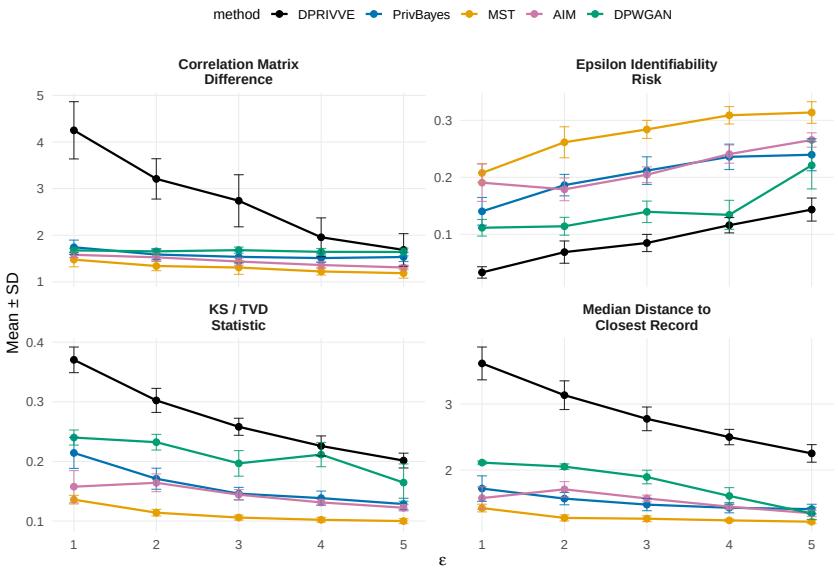


Figure 13: Additional fidelity and empirical privacy results for the DIABETES dataset across privacy budgets. Points show means and error bars show one standard deviation across synthetic-data replicates.

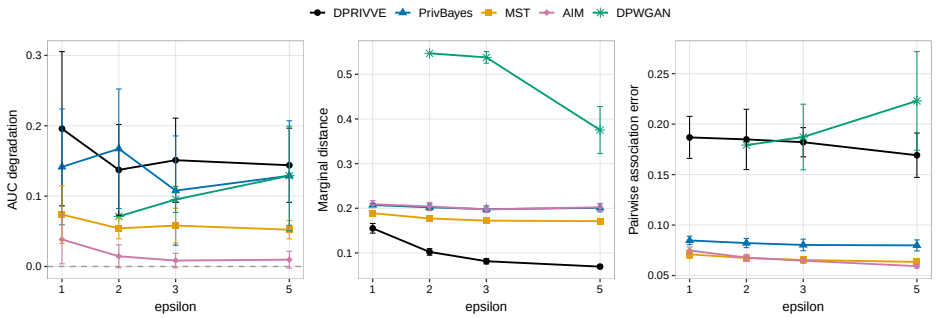


Figure 14: Additional utility and fidelity results for the FRAMINGHAM dataset across privacy budgets. Panels show degradation in downstream AUC and accuracy relative to the corresponding real-data benchmark, marginal distance, and pairwise association error. Lower values indicate better preservation. Points show means and error bars show one standard deviation across synthetic-data replicates.

A.7 Ablation Study

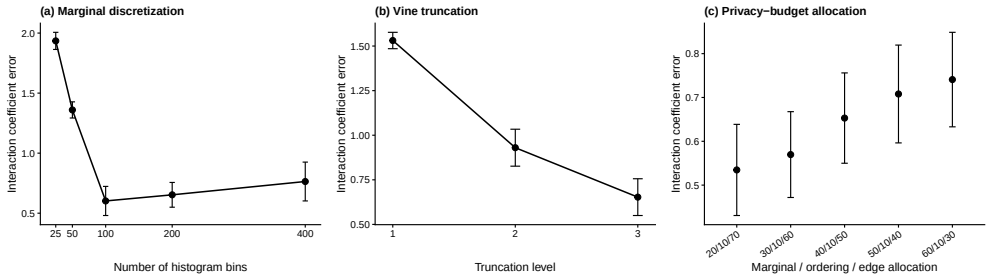


Figure 15: Ablation analysis of DPRIVVE. Panel (a) examines the number of bins used for marginal discretization, Panel (b) examines the D-vine truncation level, and Panel (c) examines the allocation of the privacy budget across marginal estimation, vine ordering, and edge learning. Points show mean interaction-coefficient error and error bars indicate variability across simulation replicates.

A.8 Additional Runtime Results

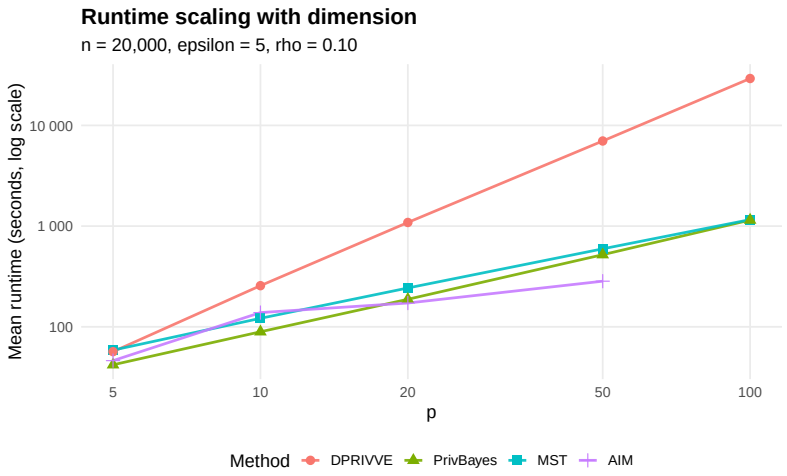


Figure 16: Runtime scaling as a function of dimensionality p.