Inferred Ergodic Dispersion to Evaluate Goodness of Synthetic Data

First Author*, Second Author† *Department, University, City, Country
Email: first.author@university.edu †Department, University, City, Country
Email: second.author@university.edu

***Abstract*—**

Introduction

Large, longitudinal data sets are required to study the causes of decline and dementia, which develop gradually over decades.[?] High-quality cognitive assessment is expensive. Thus available cohorts with high-quality cognitive measures and gold-standard tend to be relatively small.

While pooling such cohorts can improve power, it is often impossible or impractical to pool data from multiple cohorts. For example, to protect participant information, research data are increasingly housed in high-security enclaves. Synthetic data approaches – in which models fit in each data set are used to generate synthetic data, and the resulting datasets are pooled – have been proposed as an approach for conducting analyses combining information from multiple cohorts.[?] However, verifying that the generated data captures the *structural dependencies* present in the real data remains challenging.

Previous literature A previous literature review called for additional methods for evaluating the performance of synthetic data generation algorithms. - https://www.mdpi.com/2227-7390/10/15/2733 -

We propose an approach for assessing the quality of synthetic datasets and evaluate its performance in an application using real data.

-Key features of this approach and how it compares to other approaches.

Methods

Data

For our data application, we used data from wave H (2002) of the Health and Retirement Study (HRS).[?] The HRS is a nationally-representative longitudinal cohort of adults aged 50 and over and their spouses of any age, including socioeconomic, health and cognitive data.

[data approvals etc.]

Analysis

We compare Synthetic data generation 1) without perturbation vs. 2) with systematic perturbation.

We hypothesize that 1) difference between synthetic data and real data distribution is small and, 2) difference between synthetic data and real data is not 0

In the HRS data simulation, we chose ”Word recall - Delayed” as the primary outcome for comparison, the manipulating key variable is ”Educational Attainment,” dichtomized into ”High school completion or less” ($n_1 = xx$) and ”Some college or higher” ($n_2 = xx$)

Approach

We propose a general-purpose approach:

1) Train a generative model on the *real* dataset to infer the underlying dependency geometry.
2) Use the same model to measure how far both real and synthetic datasets are from their *ergodic projection* (the complete-independence baseline in LSM space).
3) Compare these dispersions to quantify the *structural fidelity* of the synthetic data.

Crucially, this requires only that a model can be trained on the real dataset—it does *not* require the model to be the generator of the synthetic data. Any generative model can be evaluated in this way.

Let us consider a system with N discrete variables $X_1, \dots, X_N$ taking values in finite alphabets $\Sigma_1, \dots, \Sigma_N$. For each i, we assume that we have a model that specifies a *component predictor*

$$\phi_i : \Sigma_{-i} \to \Delta(\Sigma_i),$$

which returns a conditional distribution over X_i given an assignment x_{-i} to all other variables.

$$\hat{x}^i = \phi_i(x_{-i})$$

Thus, the estimate $\hat{x}^{\,i}$ is produced by sampling $\hat{x}^{\,i} \sim \phi_i(x_{-i})$.

In the *ideal* (oracle) case for the original data-generating process with joint law P, we have $\phi_i(x_{-i}) = P(X_i \mid x_{-i})$.

In the LSM framework, we assume our component predictors are conditional inference trees.

LSM-induced geometry: For two samples $x, y \in \Sigma_1 \times \cdots \times \Sigma_N$, define the LSM distance

$$\theta(x, y) = \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{JS}}\big(\, \phi_i(x^{-i})\, ,\, \phi_i(y^{-i})\,\big), \qquad (1)$$

where D_{JS} is the Jensen–Shannon divergence. For a distribution R over full assignments, we will use pointwise plug-in functionals of the form $x \mapsto \theta(x, \cdot)$ and take expectations under R.

Definition 1 (Ergodic Projection)**.** *Let $\phi_i(\varnothing) \in \Delta(\Sigma_i)$ denote the unconditional marginal for X_i obtained by* removing all cross-variable dependencies*. The* ergodic projection *is the product measure*

$$\psi^\star \triangleq \bigotimes_{i=1}^{N} \phi_i(\varnothing), \qquad (2)$$

which is the maximum-entropy distribution consistent with the model-implied marginals. In particular, when $x' \sim \psi^\star$, the inputs to ϕ_i carry no information from X_{-i}.

Definition 2 (Ergodic Dispersion)**.** *The distance of a randomly chosen sample $x \in D$ from the ergodic projection defines a random variable $\delta(D)$ which we refer to as the ergodic dispersion of the data set:*

$$\delta(D) = \theta(x, \psi^\star), x \in D \qquad (3)$$

Values near 0 indicate conditional structure close to independence.

Goodness-of-Synthesis: Given the original dataset D and a synthetic version D', we generate the random variables $\delta(D), \delta(D')$. Our validation objective is to test whether the dispersion distribution induced by the synthetic data matches that of the real data:

$$H_0 : \delta(D) = \delta(D') \qquad (4)$$

In practice, we apply a two-sample distributional test to the observed sets $\{\bar{\delta}(X_m)\}_{m=1}^{M}$ and $\{\bar{\delta}(Y_n)\}_{n=1}^{N}$: (i) the Kolmogorov–Smirnov two-sample test when treating $\bar{\delta}$ as continuous; (ii) a χ^2 test on quantile-binned histograms when discreteness or ties are substantial; and, where appropriate, a permutation test on differences of means or energy-distance/MMD as robustness checks. Rejection of H_0 indicates a mismatch in dependency structure relative to the independence baseline, i.e., the synthetic generator does not faithfully reproduce the LSM-implied conditional geometry of the real data.

Comparison Criteria

Between original and synthesized data, we compare: 1) Data distribution 2) Distribution of primary outcome variable with rigorous statistical test (or regression) 3) Variance-covariance structure 4) Partial correlation matrix or network 5) Wasserstein distance between original and synthesized data

Discussion

Advantages

NEEDS TO BE EDITED

- **Model-agnostic evaluation**: Any synthetic data generator can be assessed.
- **Dependency-aware**: Captures high-order correlations beyond marginals.
- **Single scalar metric**: Easy to compare across models, datasets, and domains.
- **No task dependence**: Does not require labeled data or downstream models.
- **Interpretable baseline**: Ergodic projection is a natural, parameter-free independence reference.

Limitations/caveats

- **Causal inference** All normal causality caveats still apply: When this method succeeds, the relationships between the variables are ”the same” in the synthetic and original data. Identification of the causal effect in the synthetic dataset requires that the causal effect is identified in the original dataset, and any analysis
- **Reidentification of participants** Data use agreements may preclude sharing of synthetic data, which may be used to reidentify participants.

Conclusion

NEEDS EDITING

We present a principled method for evaluating synthetic data quality using *ergodic dispersion* from the LSM framework. By training the LSM only on real data, then comparing the distance of real and synthetic datasets from the ergodic projection, we obtain a simple, interpretable, and domain-independent measure of structural fidelity. This method applies even when the synthetic data is generated by completely different models, making it a versatile addition to the synthetic data evaluation toolkit.