Comparison of Various Data Generation Methods and Datasets

...

Ritik Dutta, IIT Gandhinagar

Summary

- Goals
- Data Generation Methods
- Datasets Considered
- Metrics Used
- Related Works
- Challenges and Future Work

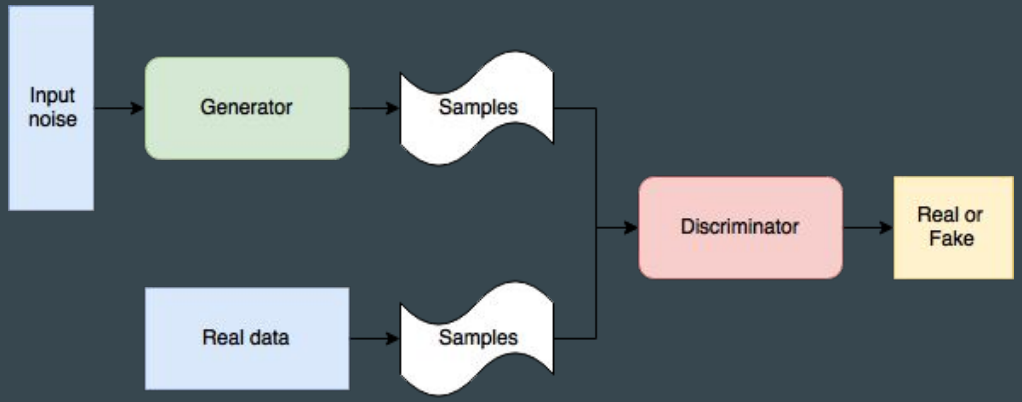
Goals

- Quality control of generated data
  - Good marginal distributions
  - Retain interdependency of features
  - Shouldn’t be copies of real data

Methods for Data Generation

- GANs
  - Wasserstein GAN
  - medGAN
- Multiple Imputations using Random Forests
- Copulas
- SAM: Structural Agnostic Model

Generative Adversarial Networks

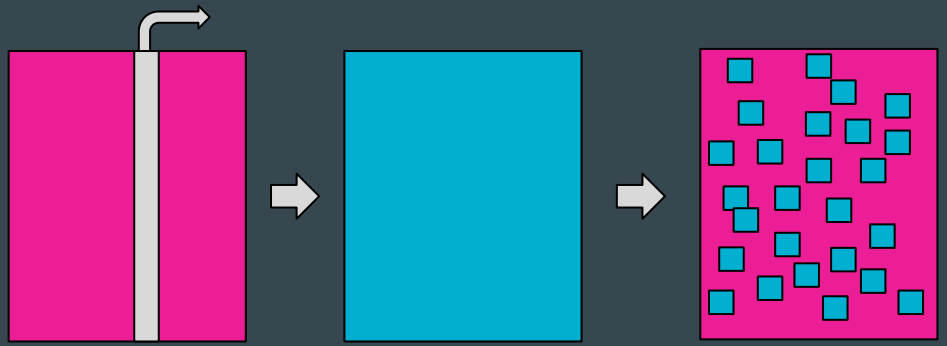


GAN Architecture

- Two versions of GANs used for data generation:
  - medGAN
  - Wasserstein GAN

Multiple Imputation using Random Forests

- Ensemble ML method using multiple decision trees
- Implementation:
  - Train RF for each feature
  - Predict values using RF
  - Replace in original matrix



ORIGINAL DATA

(1) Build predictors of one column from the others..

PREDICTED DATA

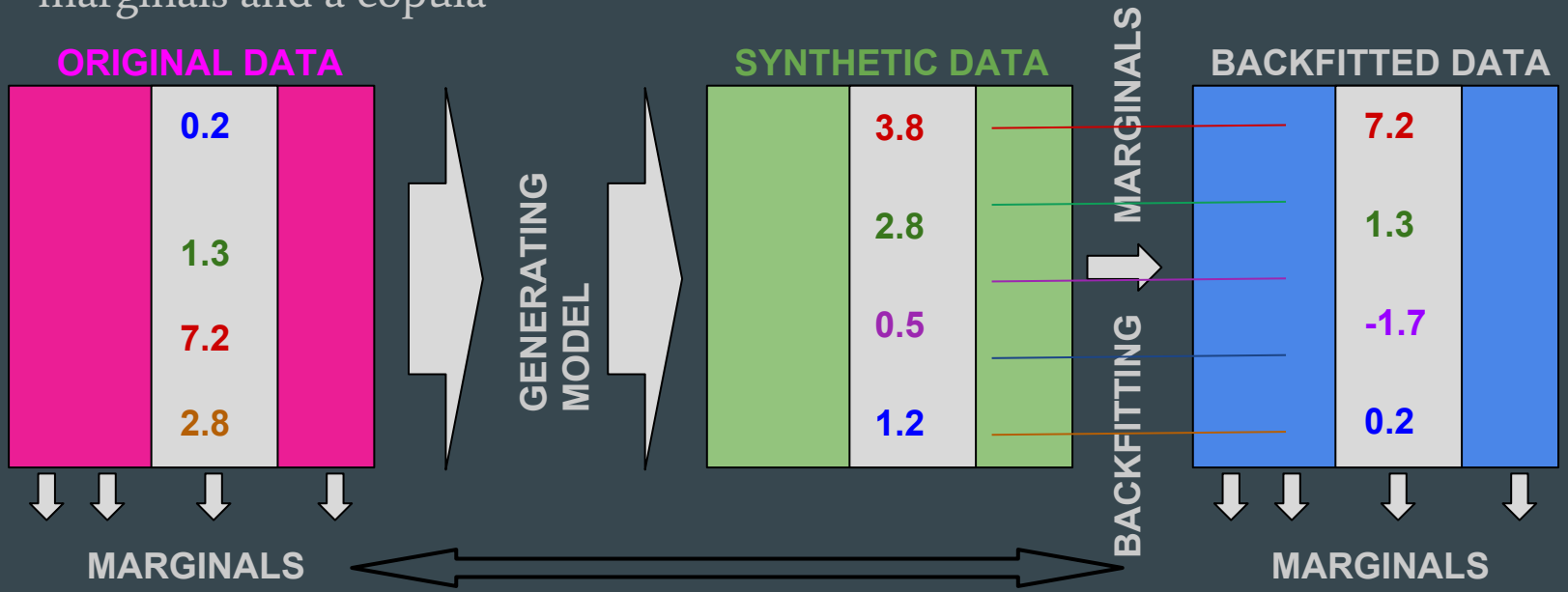
(2) Make a matrix of predictions.

SYNTHETIC DATA

(3) Mix original and predicted data randomly in a certain proportion p.

Copulas

- Copulas: Multivariate distribution with uniform marginals
- Sklar's theorem: Every multivariate distribution can be expressed in terms of its marginals and a copula



Datasets considered

Boston Housing Dataset	Adult Dataset	Iris Dataset
Concerning housing in the area of Boston. Consists variables like per capita crime rate	Consists of data pertaining to age, gender, workclass, education, to determine income	Consists information on different types of the Iris plant. Consists attributes such as petal length, petal width, etc.
Contains numeric and binary data	Contains categorical and integer data	Contains only numerical data
15 features, ~130 samples	14 features, ~48,000 samples	4 features, ~35 samples

Characteristics of a good metric

- Agree with human perceptual judgments and human rankings of models
- Ability to distinguish generated samples from real ones; discriminability
- Favor models that generate diverse samples
- Have well-defined bounds (lower, upper, and chance)
- Have low sample and computational complexity

Metrics to Check Similarity of Generated Data to Original

Univariate

- Kolmogorov-Smirnov Test

Multivariate

- Maximum Mean Discrepancy
- 1-Nearest Neighbor classifier
- Lp Distance
- Principal Component Analysis

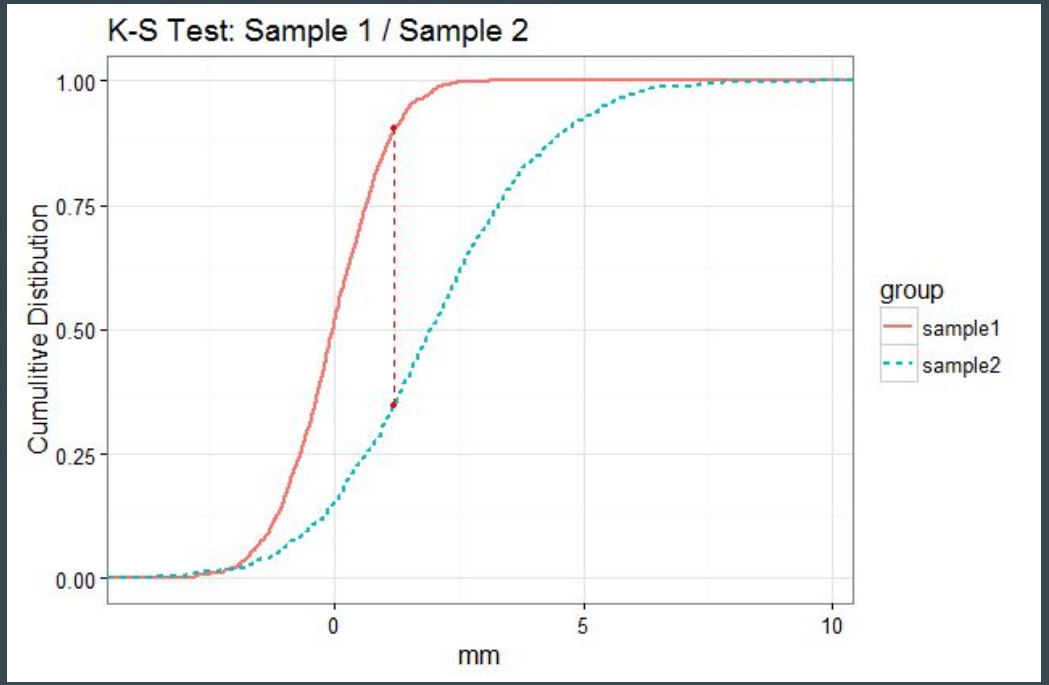
Application

- Similar covariances/ correlation
- Dimension-wise Prediction

Kolmogorov Smirnov Test

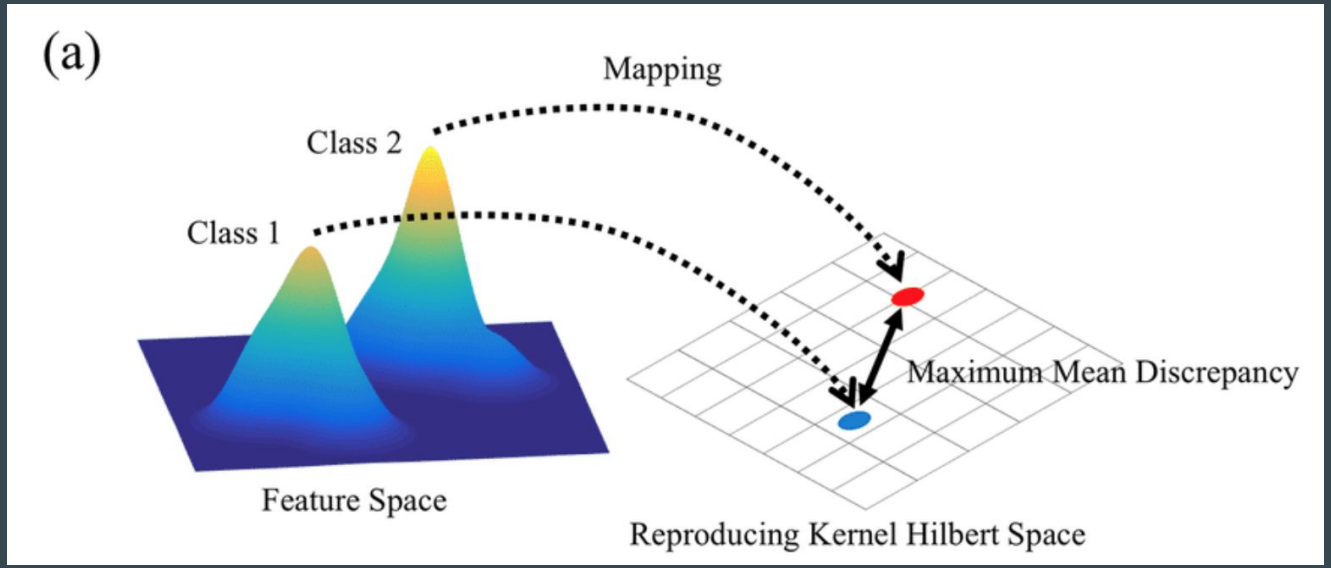
- Just the maximum absolute difference between the CDF of two populations
- If both populations come from

same distribution, the diff. should be 0

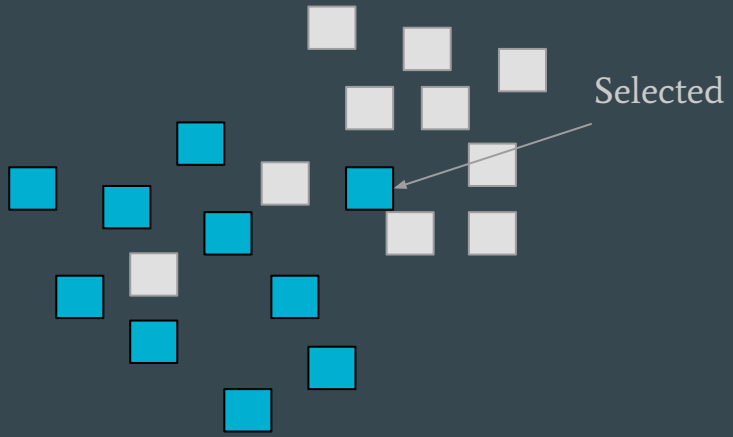


Maximum Mean Discrepancy

- Smooth function used whose value is high for points belonging to one class; low for the other class



1-Nearest Neighbour Classifier



Train classifier, and predict class of left out sample

Principal Component Analysis

- First two principal components of the two datasets can be plotted and checked if there are any significant visual differences

Dimension-wise Prediction

- Train two logistic regression models using both real and generated data leaving out one feature which is predicted
- Compare F1 score of both the models on the test set
- More similar score -> more similar distributions

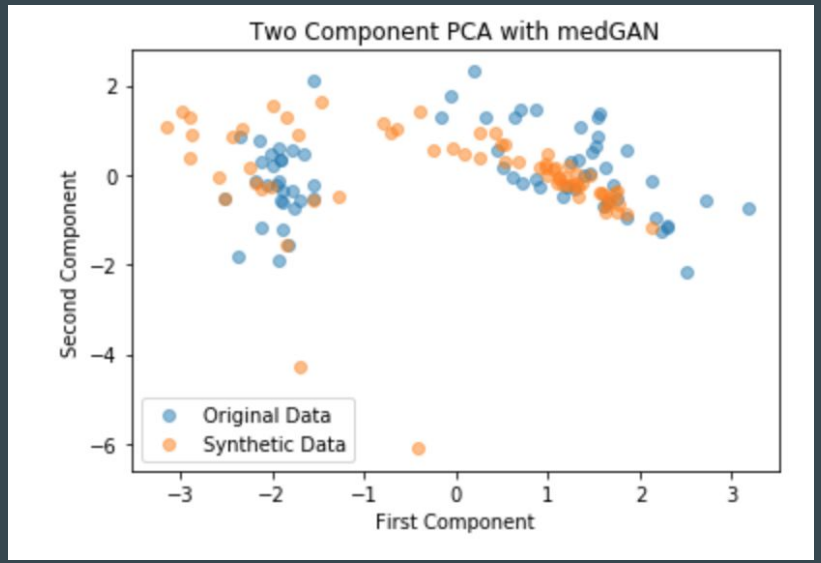
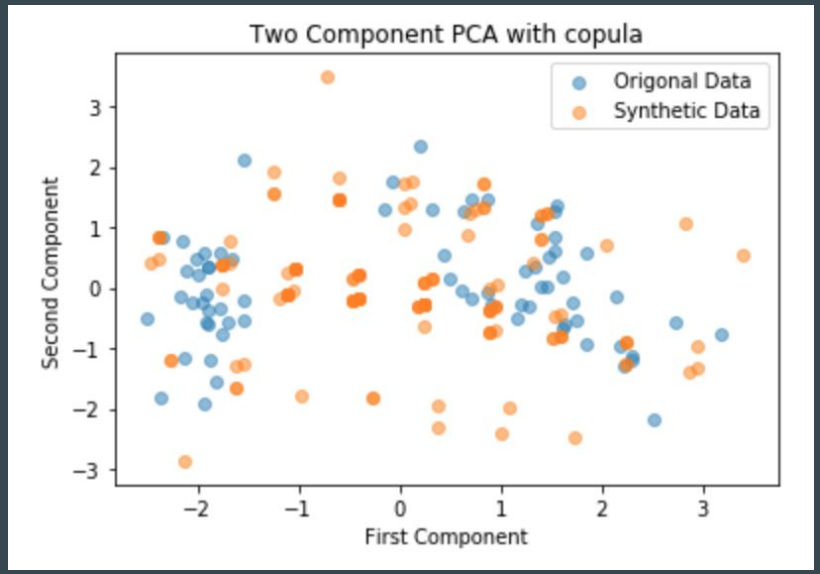
Related Works

- Daniel Jiwoong Im et al. claim that rankings produced by four metrics including 1) Jensen-Shannon Divergence, 2) Constrained Pearson χ 2 , 3) Maximum Mean Discrepancy, and 4) Wasserstein Distance, are consistent and robust across metrics.
- Some suggest that evaluation criteria should be task specific
- L. Theis et al. report that in case of image generative models, good performance in one criterion need not imply good performance in another criterion

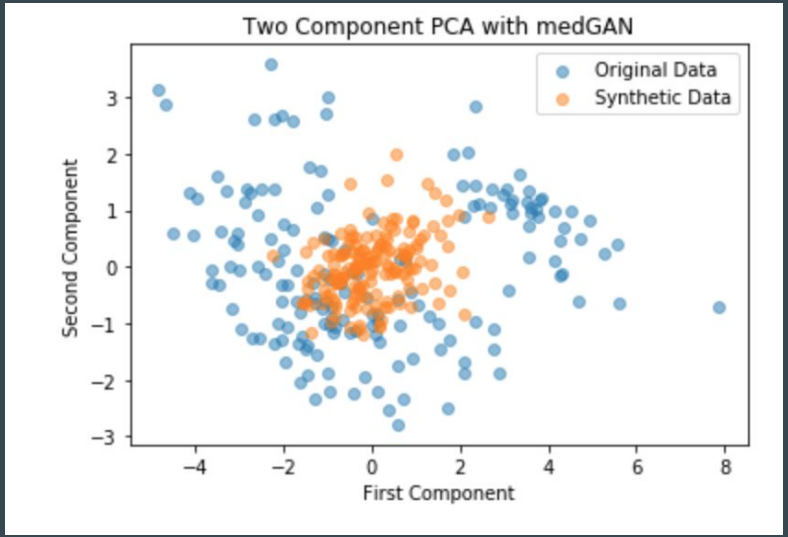
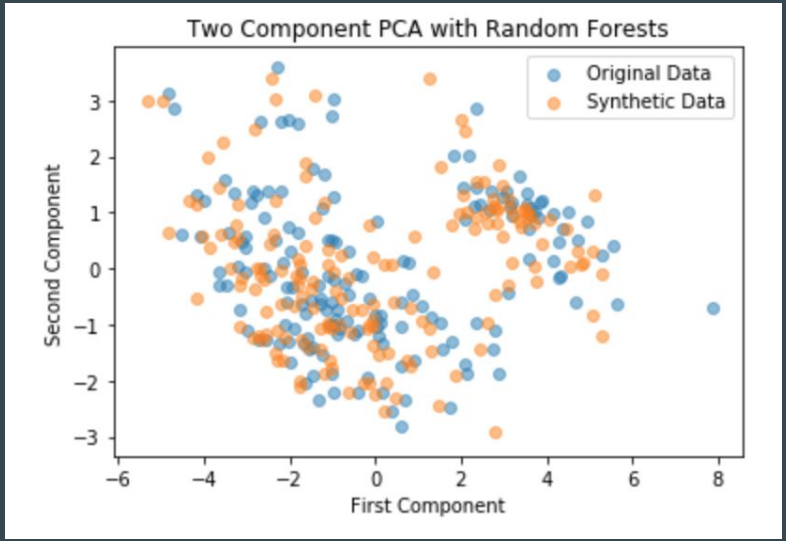
Measure		Desiderata: Discriminability	Detecting Overfitting	Disentangled Latent Spaces	Well-defined Bounds	Perceptual Judgments	Sensitivity to Distortions	Comp. & Sample Efficiency
1. Average Log- likelihood	[32, 90]	low	low	-	[-∞, ∞]	low	low	low
2. Coverage Metric	[92]	low	low	-	[0, 1]	low	low	-
3. Inception Score (IS)	[80]	high	moderate	-	[1, ∞]	high	moderate	high
4. Modified Inception Score (m-IS)	[34]	high	moderate	-	[1, ∞]	high	moderate	high
5. Mode Score (MS)	[11]	high	moderate	-	[0, ∞]	high	moderate	high
6. AM Score	[119]	high	moderate	-	[0, ∞]	high	moderate	high
7. Fréchet Inception Distance (FID)	[35]	high	moderate	-	[0, ∞]	high	high	high
8. Maximum Mean Discrepancy (MMD)	[33]	high	low	-	[0, ∞]	-	-	-
9. The Wasserstein Critic	[2]	high	moderate	-	[0, ∞]	-	-	low
10. Birthday Paradox Test	[3]	low	high	-	[1, ∞]	low	low	-
11. Classifier Two Sample Test (C2ST)	[51]	high	low	-	[0, 1]	-	-	-
12. NDB	[76]	low	high	-	[0, ∞]	-	low	-
13. Classification Performance	[73, 42]	high	low	-	[0, 1]	low	-	-
14. Image Retrieval Performance	[100]	moderate	low	-	*	low	-	-
15. Generative Adversarial Metric (GAM)	[40]	high	low	-	*	-	-	moderate
16. NRDS	[117]	high	low	-	[0, 1]	-	-	poor
17. Adversarial Accuracy & Divergence	[109]	high	low	-	[0, 1], [0, ∞]	-	-	-
18. Reconstruction Error	[107]	low	low	-	[0, ∞]	-	moderate	moderate
19. Image Quality Measures	[103, 77, 44]	low	moderate	-	*	high	high	high
20. Low-level Image Statistics	[113, 45]	low	low	-	*	low	low	-
21. Precision, Recall and F_1 score	[62]	low	high	✓	[0, 1]	-	-	-

Ali Borji. “Pros and Cons of GAN Evaluation Measures”. [1]

Iris Dataset

	medGAN	Copula
PCA		
MMD	0.148	0.508
KS Test	0.32, 0.38, 0.32, 0.31	1, 0.99, 0.96, 0.77

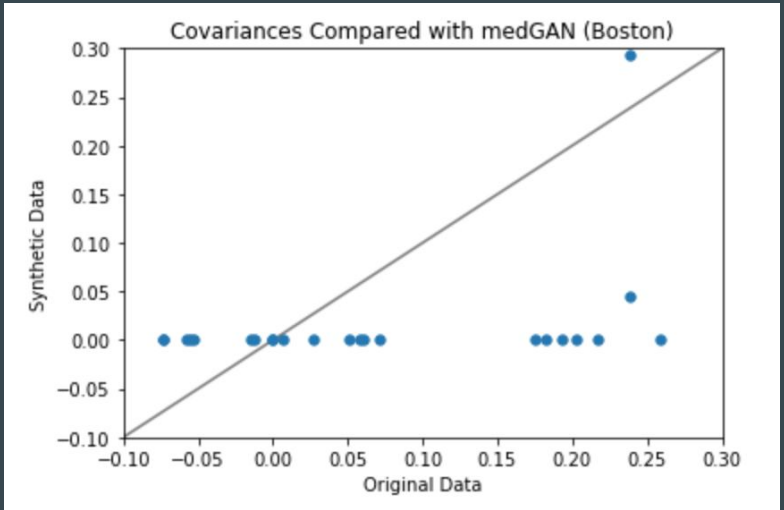
Boston Dataset

	medGAN	Imputation with RF
PCA		
MMD	0.020	0.105

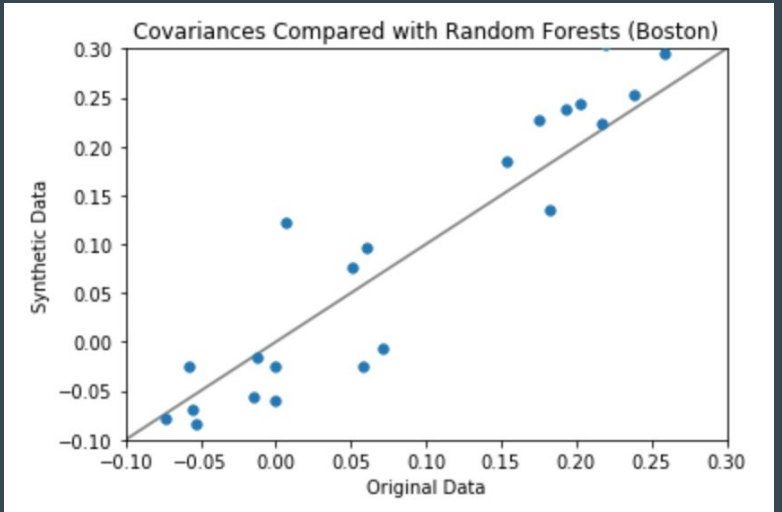
Boston Dataset

Covariance values for each feature

medGAN



Imputation with RF



Challenges and Future Work

- Discrepancy between different metrics: which metric to believe then?
- Metrics might not always work ideally: MMD might not be 0 even when Pr = Pg because of sampling variance
- Work with MIMIC, Adult, etc. on comparisons

References

[1] Ali Borji. “Pros and Cons of GAN Evaluation Measures”. In: CoRR abs/1802.03446 (2018). arXiv: 1802.03446. url: http://arxiv.org/abs/1802.03446

[2] Gao Huang et al. An empirical study on evaluation metrics of generative adversarial networks. 2018. url: https://openreview.net/forum?id=Sy1f0e-R-

[3] Daniel Jiwoong Im et al. “Quantitatively Evaluating GANs With Divergences Proposed for Training”. In: CoRR abs/1803.01045 (2018). arXiv: 1803.01045. url: http://arxiv.org/abs/1803. 01045.

[4] L. Theis, A. van den Oord, and M. Bethge. “A note on the evaluation of generative models”. In: ArXiv e-prints (Nov. 2015). arXiv: 1511.01844 [stat.ML].

[5] Discussions with Dr. Guyon