# A Hybrid Bayesian Model to Analyse Healthcare Data

Narges Pourshahrokhi*, Samaneh Kouchaki[1], Kord M. Kober[2], Christine Miaskowski[2], and Payam Barnaghi [2]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[2]School of Nursing, University of California, San Francisco
[3]Department of Brain Sciences, Imperial College London

## Abstract

Missing values exist in nearly all clinical studies because data for a variable or question are not collected or not available. Imputing missing values and augmenting data can significantly improve generalisation and avoid bias in machine learning models. We propose a Hybrid Bayesian inference using Hamiltonian Monte Carlo (F-HMC) as a more practical approach to process cross-dimensional relations by applying a random walk and Hamiltonian dynamics to adapt posterior distribution and generate large-scale samples. The proposed method is applied to cancer symptom assessment, and MNIST datasets confirmed to enrich data quality in precision, accuracy, recall, F1-score, and propensity metric.

## 1 Intoduction

Many large datasets are inherently uncertain due to noise, incompleteness, inconsistency, and lack of a sufficient number of training samples, which significantly impact the outcomes of the machine learning techniques by misleading or biasing the final results. Using augmentation and imputation [3] methods can improve data quality. However, they usually suffer from bias caused by dropping cases or replacing data with seemingly suitable values. Bayesian inference with Hamiltonian Monte Carlo offers a very efficient way to process high-dimensional and small sample datasets. In this paper, we apply our proposed model to the MNIST dataset and a dataset collected from 1342 cancer patients symptoms during chemotherapy by a team in the School of the Nursing University of California (USCF)[1].

## 2 Related Work

Sampler methods such as Gibbs and MCMC to estimate parameters of interest under missing values are the closest techniques to our proposed model [4, 2]. They are based on setting a fixed parameter and require more computation time, particularly for large sample size studies. The proposed approach opens a new window of using samplers to estimate missing values directly instead of estimating the model parameters. In particular, this method will enrich the sampler to explore high dimensional data for estimating missing values while the generated samples preserve data privacy. More details of the model are presented in the Method section.

## 3 Method

The Bayesian approach provides a framework for making inferences with incomplete data by considering the full-data model as the posterior. Let $Y = (Y_{ij})$ denote a rectangular data set where $i$

*Corresponding author email: n.pourshahrokhi@surrey.ac.uk

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

is the data sample, and $j$ is variables' features. Let's partition Y into observed and missing values, $Y = (Y^{obs}, Y^{mis})$. Let $M$ be as the mask vector which indicates observed components in $Y_{ij}$ is defined as:

$$M_{ij} = \begin{cases} 1, & Y_{ij}\ missing\ . \\ 0, & Y_{ij}\ observed. \end{cases} \tag{1}$$

One can specify the full-data response by calculating the joint model where $w$ is an unknown parameter which consists of $\theta$ and $\phi$. Then the joint model (likelihood) of the full data is

$$p(Y, M|\theta, \phi) = p(Y^{obs}, Y^{mis}, M|\theta, \phi) \tag{2}$$

The joint model in Eq. 2 cannot be evaluated in the usual way because it depends on missing data. However, the marginal distribution of the observed data can be obtained by integrating out the missing data. Consequently, the joint model can be written as follows after applying the conditional independence assumption and selection model factorisation:

$$p(Y^{obs}, M|\theta, \phi) = \int p(Y^{obs}, Y^{mis}, M|\theta, \phi) dy^{mis} = \int p(M|Y^{obs}, Y^{mis}, \theta) p(Y^{obs}, Y^{mis}|\phi) dy^{mis} \tag{3}$$

By estimating the integral in Eq. 3, one can determine full data response and consequently generate new samples to utilise for data imputation and augmentation. We show that Monte Carlo samplers, especially F-HMC samplers, are an effective method for the Eq. 3 estimation in high-dimensional data.

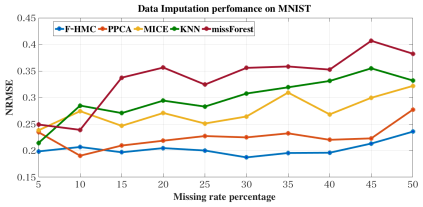
### 3.1 F-HMC Algorithm for Imputation and Augmentation

We consider the problem formulation as a *d*-dimensional space that X = $(X_1, ..., X_d)$ is a random variable selected from it. We consider $M = (M_1, ..., M_d)$ as a mask vector which identifies the missing values in the dataset D as defined in the equation 1. We also define dataset $D = \{(x^i, m^i)\}$, where $m^i$ is the obtained realisation of $M$ corresponding to $x_i$. Our goal is to impute the unobserved values in each $x^i$.

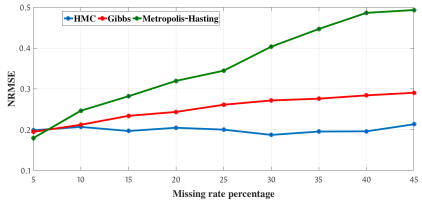
Given a dataset $D$ and a mask vector $M$, each missing data considered as an unknown parameter which their possible values are drawn and directed to $\bigtriangledown(-\log(P(X|D))$, i.e. the potential energy in Hamiltonian Monte Carlo semantic. HMC first models the posterior distribution of each feature dimension of data using the Gaussian likelihoods with a Laplacian prior, to find the $mu$ and $sigma$s of feature distribution. Then, all the $mu$ and $sigma$s for each feature dimension are given to another HMC (fold) with respect to the cross-correlation between all features. The F-HMC adopts the results using gradient information to draw samples from the cross-dimensional distribution of features. After a burn-in time, the algorithm converges, and it can generate samples that belong to the posterior of the complete dataset. For imputation, the missing values are replaced by marginalisation over generated samples correspond to that missing part. Algorithm 3.1 presents the steps in our approach, where $D$ is the incomplete dataset, $M$ is the mask vector to indicate missing values in $D$, $\eta$ is the step size for HMC dynamics, and $k$ is the number of generated samples of the full dataset.

**Input** : $D, M, k, \eta$
**Output** : $X^1, X^2, X^3$,... , $X^K$
**Initialisation** : $X^0$
**for** $i \leftarrow 1$ to $k$ **do**
  $X_0 \leftarrow X^{k-1}$
  $mu_d, sigma_d \leftarrow HMC$ for each
  $x_1, x_2, \ldots, x_d$ separately;
  $MU, SIG \leftarrow$ joint $mu_d, sigma_d$
  $X_{k-1} \leftarrow HMC(mu_d, sigma_d, \eta)$
  $X_k \leftarrow X_{k-1} \oplus M$
  $X^k \leftarrow X_k$
**end for**

After initialisation of $X_0$ with white noise, the algorithm starts. First HMC iterates in parallel over each feature dimension of $D$ to calculate $mu_d, sigma_d$, separately. Next, it produces new $MU$ and $SIG$ by concatenating the $mu_d, sigma_d$ which allows the cross-correlation of features to be considered in the learning. Then $MU$ and $SIG$ are given to the second HMC sampler to explore




(a) Comparison of various data imputation techniques; KNN, MICE, missForest, and the F-HMC on MNIST dataset




(b) Comparing the impact of sampler's type on data imputation performance using the proposed approach on MNIST

Figure 1: NRMSE results on the MNIST. the lower score shows higher data quality

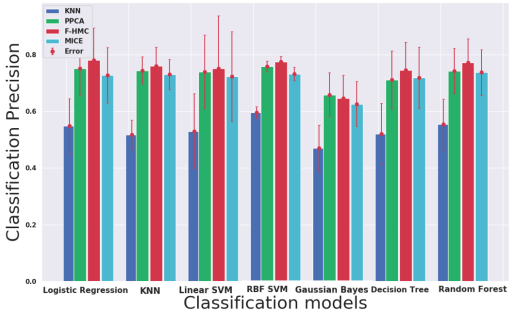



Figure 2: Classification performance of various machine learning techniques on an imputed version of the USCF data after applying KNN, PPCA, F-HMC, MICE. A higher score indicates a more reliable imputation method.

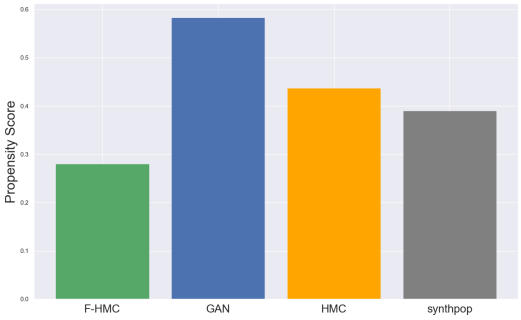



Figure 3: Comparing propensity score in data augmentation on the USCF dataset as a metric to evaluate the quality of synthesised data. A lower score indicates higher quality.

the posterior and return samples from that posterior. Operator $\oplus$ is for marginalising generated samples over missing values with the help of matrix $M$. The algorithm keeps a record of the latest state of outputs for the next round to make it more accurate over time. The outputs of the algorithm $X^1, X^2, \ldots, X^k$ are the complete datasets drawn from the estimated posterior. The algorithm can impute missing values and generate more samples from posterior, in case of requiring more samples from the data.

## 4 Experiments and Evaluations

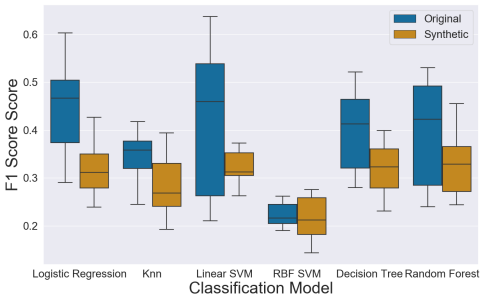
To evaluate the proposed method's performance in data imputation and augmentation, we have intended three measurement levels: **Distance metric** such as Normalised Root Mean Square Error (NRMSE) (suitable for scenarios where missing values are dropped randomly and true values exist). **Outcome performances** on classification (where we do not grand truth to indicate the actual value of missing data). **Propensity** metrics as a quality measure of augmentation performance.

### 4.1 MNIST Data

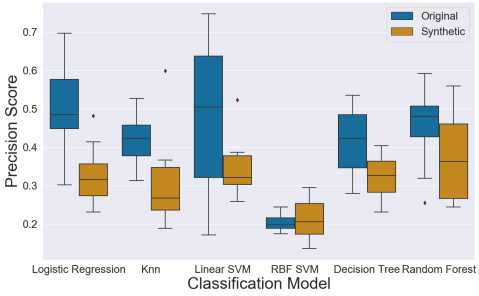
MNIST dataset containing 60k images of handwritten digits of 28*28 pixels is chosen to evaluate we consider each pixel as a feature dimension, and missing pixels are dropped randomly and reconstructed using our proposed and baseline methods. The proposed model outperforms in terms of NRMSE as shown in figure 1a. Comparisons on using various samplers in our proposed model is shown in Figure 1b .

### 4.2 The Cancer Symptom Management Dataset (USCF Dataset)

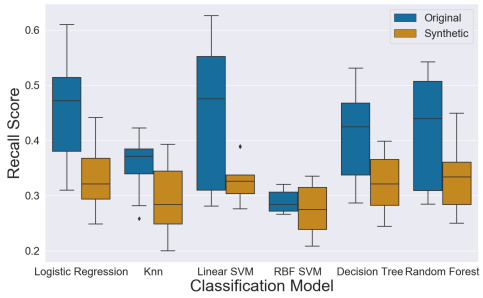
In the UCSF dataset, we are not aware of the value of missing parts, so the evaluation is done based on the classification performance after applying imputation and augmentation techniques. More improvement over classification stands for higher enrichment in data quality, hence a more reliable imputation technique. Figure 2 shows that the predicted precision in the data imputed by the proposed



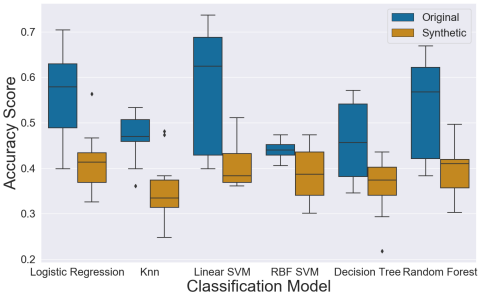

(a) F1-score




(b) Precision




(a) Recall




(b) Accuracy

Figure 5: Reproducibility of classification performance on original and synthetic USCF data using the proposed approach after applying various classification models. For instance, Logistic Regression has higher accuracy than nearest neighbour in both original and augmented data. Kernel SVM recall is less than recall in the decision tree on original data that agrees with the same results on synthetic data

approach is higher than the baseline methods and confirms that imputation using the F-HMC is more substantial than baselines. The propensity score is the probability of a given data point being assigned to a particular class. As Figure 3 shows, the synthetic data generated by the F-HMC approach has the lowest propensity score meaning the higher quality of data. It is expected that certain ML methods perform similarly on both original and augmented data considering uncertainty. We judged the performance based on the accuracy, precision, recall, and F1 score of data classification on each cancer type reported in macro-averaging as shown in Figure 5.

## 5 Conclusion

This work proposes a Hybrid Bayesian inference using Hamiltonian Monte Carlo (F-HMC) to impute missing values and generate augmented samples in high dimensional but small healthcare datasets. We demonstrate that the proposed method effectively augment data samples and impute missing values in terms of Metric distance (NRMSE), classification score impact (accuracy, recall, precision and precision F1-score) and propensity score on MNIST and a clinical dataset on cancer symptom management.

## References

[1] Papachristou et al. Congruence between latent class and k-modes analyses in the identification of oncology patients with distinct symptom experiences. *J. of Pain & Sympt. Mngm.*, 55(2), 2018.

[2] Y. Park et al. Perturbed gibbs samplers for generating large- scale privacy-safe synthetic health data. *in Proc. of IEEE Int. Conf. on Healthcare Informatics*, 2013.

[3] D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976.

[4] R. et al. Wei. Gsimp: A gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput Biol*, 14, 2018.