A GAUSSIAN LATENT VARIABLE MODEL FOR INCOMPLETE MIXED TYPE DATA

Marzieh Ajirak and Petar M. Djurić

Department of Electrical and Computer Engineering
Stony Brook University
Stony Brook, NY 11794

ABSTRACT

In many machine learning problems, one has to work with data of different types, including continuous, discrete, and categorical data. Further, it is often the case that many of these data are missing from the database. This paper proposes a Gaussian process framework that efficiently captures the information from mixed numerical and categorical data that effectively incorporates missing variables. First, we propose a generative model for the mixed-type data. The generative model exploits Gaussian processes with kernels constructed from the latent vectors. We also propose a method for inference of the unknowns, and in its implementation, we rely on a sparse spectrum approximation of the Gaussian processes and variational inference. We demonstrate the performance of the method for both supervised and unsupervised tasks. First, we investigate the imputation of missing variables in an unsupervised setting, and then we show the results of joint imputation and classification on IBM employee data.

***Index Terms*—** Gaussian process, incomplete data, heterogeneous data

1. INTRODUCTION

Real-world datasets are often heterogeneous and incomplete, meaning they contain different types of data that range from continuous to ordinal and categorical, with missing values at random locations. For example, electronic health records of hospitals might contain different clinical measurements, diagnoses, and demographic information about their patients [1]. They do not only contain different numerical lab values but also variables like race and blood type that are categorical. Moreover, the two types are often with missing values for various reasons.

Gaussian processes (GPs) are Bayesian models that offer distributions over functions. They are robust to over-fitting and they provide a principled way to tune hyper-parameters and uncertainty bounds over the outputs [2]. These properties are critical for tasks including non-linear function regression and density estimation [3]. The GP latent variable model [4] made it possible for GPs to be used in unsupervised cases for latent structure discovery. They have been recently proven to be expressive unsupervised methods, able to capture latent structures of complex high-dimensional data [5, 6]. However, their performance in the case of heterogenous data is poorly explored.

Naively applying GPs to mixed-type data can lead to misleading results. This is because when dealing with heterogeneous data, we need to use different types of likelihoods. For example, Gaussian likelihood for real-valued variables and Bernoulli likelihood for binary variables is required, respectively [7]. In this case, each likelihood contributes to the training objective in different ways leading to some data dimensions being poorly modeled in favor of other dimensions [8].

Building and modifying GPs that can accurately capture the information, and therefore the underlying latent structure, of mixed and incomplete datasets, may allow us to understand the data better, estimate missing values, and make predictions on unseen attributes more accurately. There have been few attempts to study mixed data with GPs [9]. More recently, [10] extends the use of supervised multi-output GPs, assuming that each output has its own likelihood function. By contrast, in this paper, we present a GP framework that efficiently extracts information from mixed data and effectively incorporates incomplete data.

Our contributions are as follows: we propose

- GP latent variable models that can handle mixed numerical and categorical data;
- missing data imputation using a trained GP latent variable model;
- joint classification and missing variables imputation in incomplete heterogeneous data.

The resulting model can be used both in unsupervised and supervised settings, including density estimation, missing data imputation, and classification of incomplete data.

2. PROBLEM STATEMENT

We consider a generative model for data represented by a matrix $\mathbf{Y}$. The rows of this matrix, $\mathbf{y}_n^\top$, $n = 1 : N$, are vectors composed of elements of two types, and n is an index that refers to a particular subject/object. The first D_c elements of $\mathbf{y}_n$, $y_{n,d}, d = 1 : D_c$, are categorical variables, and they have assigned indices from 1 to K, whereas the last D_q elements $y_{n,d}$, $d = D_c{+}1 : D_c{+}D_q$, are numerical with support on the real line. Below, we present a generative model that allows us to treat various types of problems based on this type of mixed data in a unified way.

Similar to [11, 12], we assume that each categorical variable is drawn from a corresponding probability mass function (pmf) with masses $f_{n,d,k} = \mathcal{F}_{d,k}(\mathbf{x}_n)$, $d = 1, 2, \ldots, D_c$, $k = 1, 2, \ldots, K$, where $\mathbf{x}_n \in \mathbb{R}^Q$ is an input latent variable whose prior is a zero mean Gaussian probability density function (pdf) with a given covariance matrix $\mathbf{\Sigma}$. Further, we assume that $\mathcal{F}_{d,k}$ is a function whose prior is a GP.

Similarly, each numerical variable is assumed to be a function of the input $\mathbf{x}_n$. More specifically, $y_{n,d}$ is Gaussian distributed with mean $f_{n,d}$ and variance σ_q^2. The mean $f_{n,d} = \mathcal{F}_d(\mathbf{x}_n)$, $d = D_c{+}1 : D_c{+}D_q$, is a function of $\mathbf{x}_n$ too. Again, we assume that the

The authors thank the support of NSF under Grant 2212506.

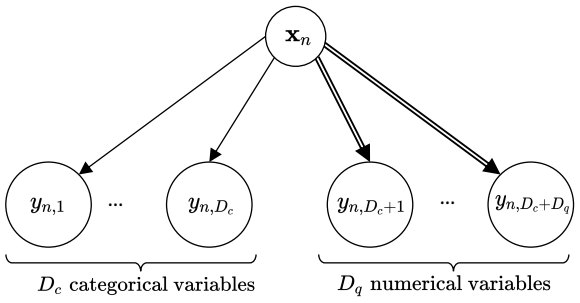

Fig. 1. Generative model, where every dimension in the observation vector $\mathbf{y}_n = [y_{n1}, \ldots, y_{nD}]$ corresponds to an either numerical or categorical variable.

functions $\mathcal{F}_d$ have GP priors. Figure 1 depicts the observation vector $\mathbf{y}_n$ and the latent variable $\mathbf{x}_n$.

The generative model can be summarized as follows:

$$\mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{\Sigma}), \quad n = 1:N, \tag{1}$$

$$\mathcal{F}_{d,k} \overset{\text{iid}}{\sim} \mathcal{GP}(0, \mathbf{K}_d), \quad d = 1:D_c, \; k = 1:K, \tag{2}$$

$$f_{n,d,k} = \mathcal{F}_{d,k}(\mathbf{x}_n), \quad d = 1:D_c, \tag{3}$$

$$\mathcal{F}_d \overset{\text{iid}}{\sim} \mathcal{GP}(0, \mathbf{K}_d), \quad d = D_c{+}1:D_c{+}D_q, \tag{4}$$

$$f_{n,d} = \mathcal{F}_d(\mathbf{x}_n), \; d = D_c{+}1:D_c{+}D_q, \tag{5}$$

$$p(y_{n,d} = k) = \frac{\exp(f_{n,d,k})}{\sum_{k'=1}^{K} \exp(f_{n,d,k'})}, \quad d = 1:D_c, \tag{6}$$

$$p(y_{n,d}) = \mathcal{N}(f_{n,d}, \sigma_q^2), \quad d = D_c{+}1:D_c{+}D_q. \tag{7}$$

In words, (1) suggests that the latent vectors $\mathbf{x}_n \in \mathbb{R}^Q$, $n = 1, 2, \ldots, N$, are independently generated from a multivariate zero mean Gaussian pdf with a covariance matrix $\mathbf{\Sigma}$. The $D_c \times K$ functions $\mathcal{F}_{d,k}$, $d = 1:D_c, k = 1:K$ have as priors zero mean GPs (denoted by $\mathcal{GP}$) and where for each dimension, the GPs may have different covariance matrices $\mathbf{K}_d$ (cf. (2)). The values of the weights are obtained via (3), where the weight that corresponds to the dth categorical variable of the nth subject/object and its kth value is a function of $\mathbf{x}_n$. The notation $\mathcal{F}_{d,k} \overset{\text{iid}}{\sim} \mathcal{GP}(0, \mathbf{K}_d)$ means that any finite collection of variables $f_{n,d,k}$ for given d and k are jointly Gaussian. with mean zero and covariance $\text{cov}(f(\mathbf{x}_n)f(\mathbf{x}_{n'})))$. Similarly, as in (2), the D_q functions $\mathcal{F}_d$, $d = D_c + 1 : D_c + D_d$ have priors from the same zero mean GPs with a covariance matrix $\mathbf{K}_d$ (cf. (4)). The means of the Gaussians used for generating the numerical values $y_{n,d}$, $f_{n,d}$, are obtained by (5), and clearly, they depend on $\mathbf{x}_n$. The probability masses of the categorical variables are obtained by the softmax function (6), and the pdf of the numerical variable $y_{n,d}$ is Gaussian as shown by (7). In summary, first we generate N vectors $\mathbf{x}_n$, where $n = 1, 2, \ldots, N$. Given all the $\mathbf{x}_n$s, for each d we construct a covariance matrix $\mathbf{K}_d \in \mathbb{R}^{N \times N}$ that will be used for generating the vectors $\mathbf{f}_{n,d,k} \in \mathbb{R}^{N \times 1}$, $d = 1 : D_c$ and $\mathbf{f}_d \in \mathbb{R}^{N \times 1}$, $d = D_c + 1, D_c + D_q$, using zero mean Gaussian pdfs with respective covariance matrices $\mathbf{K}_d$. From the obtained $\mathbf{f}_{n,d,k}$, we compute the pmfs of the categorical variables for each n by (6), and then draw the $y_{n,d}$ categorical variables ($d = 1 : D_c$) from the respectively obtained pmfs. We generate the numerical variables $y_{n,d}$ ($d = D_c + 1 : D_c + D_q$) from the Gaussians in (7).

In our inference, we will rely on the marginal log-likelihood of the data. Finding the marginal log-likelihood is intractable because of 1) the covariance function of the GP and 2) the softmax likelihood. To tackle these two challenges, we use sparse spectrum approximation of GPs to deal with their inverse kernel matrices and variational inference. Before we present our proposed solution, we provide some background on sparse spectrum approximation of GPs and variational inference. We also discuss the reasons why we adopt them in our proposed approach.

3. BACKGROUND

3.1. Sparse Spectrum Approximation of Gaussian Processes

We use Bochner's theorem to reformulate the covariance function in terms of its frequencies [13]. Our covariance function $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ is stationary, and thus, it can be represented as $\mathbf{K}(\mathbf{x}-\mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^Q$. According to the theorem, $\mathbf{K}(\mathbf{x} - \mathbf{x}')$ can be represented as the Fourier transform of some finite measure $\sigma^2 p(\boldsymbol{\omega})$ where $p(\boldsymbol{\omega})$ is proportional to the power spectral density of the kernel, i.e.,

$$\begin{aligned}\mathbf{K}(\mathbf{x} - \mathbf{x}') &= \int_{\mathbb{R}^Q} \sigma^2 p(\boldsymbol{\omega}) e^{-2\pi i \boldsymbol{\omega}^T(\mathbf{x}-\mathbf{x}')} \mathrm{d}\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^Q} \sigma^2 p(\boldsymbol{\omega}) \cos\left(2\pi \boldsymbol{\omega}^T(\mathbf{x} - \mathbf{x}')\right) \mathrm{d}\boldsymbol{\omega}, \end{aligned} \tag{8}$$

where $i = \sqrt{-1}$. The second equality holds because the covariance function is real-valued. The above integration can be approximately computed by the Monte Carlo method as a finite sum with J terms according to

$$\mathbf{K}(\mathbf{x} - \mathbf{x}') \approx \frac{\sigma^2}{J} \sum_{j=1}^{J} \cos\left(2\pi \boldsymbol{\omega}_j^T \left((\mathbf{x} - \mathbf{z}_j) - (\mathbf{x}' - \mathbf{z}_j)\right)\right) \tag{9}$$

with $\boldsymbol{\omega}_j \sim p(\boldsymbol{\omega})$ and $\mathbf{z}_j$ being Q dimensional vectors for $j = 1 : J$, and where they act as inducing inputs. We rewrite the above terms for every j as

$$\begin{aligned}&\cos\left(2\pi \boldsymbol{\omega}_j^T \left((\mathbf{x} - \mathbf{z}_j) - (\mathbf{x}' - \mathbf{z}_j)\right)\right) \\ &\quad = \int_0^{2\pi} \frac{1}{2\pi} \sqrt{2} \cos\left(2\pi \boldsymbol{\omega}_j^T (\mathbf{x} - \mathbf{z}_j) + \varphi\right) \\ &\quad \times \sqrt{2} \cos\left(2\pi \boldsymbol{\omega}_j^T (\mathbf{x}' - \mathbf{z}_j) + \varphi\right) \mathrm{d}\varphi. \end{aligned} \tag{10}$$

This integral can again be approximated as a finite sum using Monte Carlo integration. To keep the computations low, we approximate the integral with a single sample for every j. Then we write,

$$\begin{aligned}\mathbf{K}(\mathbf{x} - \mathbf{x}') \approx& \frac{\sigma^2}{J} \sum_{j=1}^{J} \sqrt{2} \cos\left(2\pi \boldsymbol{\omega}_j^T (\mathbf{x} - \mathbf{z}_j) + \varphi_j\right) \\ &\times \sqrt{2} \cos\left(2\pi \boldsymbol{\omega}_k^T (\mathbf{x}' - \mathbf{z}_j) + \varphi_j\right) \\ =& \widehat{\mathbf{K}}(\mathbf{x} - \mathbf{x}'), \end{aligned} \tag{11}$$

where $\varphi_j \sim \text{Unif}[0, 2\pi]$. In summary, with (11) we define our approximation of the covariance function $\widehat{\mathbf{K}}$.

We refer to $(\boldsymbol{\omega}_j)_{j=1}^J$ as inducing frequencies and to $(\varphi_j)_{j=1}^J$ as phases, and we denote $\mathbf{w} = (\boldsymbol{\omega}_j, \varphi_j)_{j=1}^J$. If we use $\widehat{\mathbf{K}}$ as the covariance function of the GP, we obtain the following generative model:

$$\boldsymbol{\omega}_j \sim p(\boldsymbol{\omega}), \quad \varphi_j \sim \text{Unif}[0, 2\pi], \quad j = 1 : J, \tag{12}$$

$$\mathbf{w} = (\boldsymbol{\omega}_j, \varphi_j)_{j=1}^J. \tag{13}$$

Clearly, we can condition this model on the finite set of random variables $\mathbf{w}$. With our modeling assumption, the model depends on these variables alone, making them sufficient statistics for the model.

3.2. Variational Inference

The predictive distribution for a new test point x^* is given by

$$p\left(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right)=\int p\left(\mathbf{y}^*|\mathbf{x}^*,\mathbf{w}\right)p(\mathbf{w}|\mathbf{X},\mathbf{Y})\mathrm{d}\mathbf{w}, \tag{14}$$

where $\mathbf{y}^* \in \mathbb{R}^D$. The distribution $p(\mathbf{w}|\mathbf{X},\mathbf{Y})$ cannot be usually evaluated analytically. Instead, an approximating variational distribution $q(\mathbf{w})$ whose structure is easy to evaluate is defined. We want our approximation distribution to be close to the posterior distribution. We, therefore, minimize the Kullback-Leibler (KL) divergence, a measure of similarity between two distributions:

$$\mathrm{KL}(q(\mathbf{w})\|p(\mathbf{w}|\mathbf{X},\mathbf{Y})) \tag{15}$$

resulting in the approximate predictive distribution

$$q\left(\mathbf{y}^*|\mathbf{x}^*\right)=\int p\left(\mathbf{y}^*|\mathbf{x}^*,\mathbf{w}\right)q(\mathbf{w})\mathrm{d}\mathbf{w}. \tag{16}$$

Minimizing the KL divergence is equivalent to maximizing the log evidence lower bound (ELBO)

$$\mathcal{L}_{\mathrm{VI}}:=\int q(\mathbf{w})\log p(\mathbf{Y}|\mathbf{X},\mathbf{w})\mathrm{d}\mathbf{w}-\mathrm{KL}(q(\mathbf{w})\|p(\mathbf{w})) \tag{17}$$

with respect to the variational parameters that define $q(\mathbf{w})$. We note that the KL divergence in the last equation is between the approximate posterior and the true posterior over $\mathbf{w}$. Maximizing this objective will result in a variational distribution $q(\mathbf{w})$ that explains the data well while still being close to the prior and preventing the model from over-fitting.

4. PROPOSED SOLUTION

Let $\mathcal{O}_n$ be the index set of observed and $\mathcal{M}_n$ the index set of missing attributes of the nth vector $\mathbf{y}_n$. Also, let $\mathbf{y}_n^o$ be a vector of observed values (the indices of its elements come from $\mathcal{O}_n$), and $\mathbf{y}_n^m$ the vector of missing values (the indices of its elements come from $\mathcal{M}_n$). For the joint distribution of $\mathbf{y}_n$ and $\mathbf{x}_n$ we write

$$p\left(\mathbf{y}_n,\mathbf{x}_n\right)=p\left(\mathbf{x}_n\right)\prod_d p\left(y_{nd}|\mathbf{x}_n\right). \tag{18}$$

The above factorization of the likelihood of the unknowns allows us to separate the contributions of the observed data $\mathbf{y}_n^o$ from those of the missing data $\mathbf{y}_n^m$. We have

$$p\left(\mathbf{y}_n|\mathbf{x}_n\right)=\prod_{d\in\mathcal{O}_n}p\left(y_{n,d}|\mathbf{x}_n\right)\prod_{d\in\mathcal{M}_n}p\left(y_{nd}|\mathbf{x}_n\right). \tag{19}$$

The ELBO of the marginal likelihood of observed data, $\mathbf{Y}^o$, can be written as

$$\begin{aligned}\log p(\mathbf{Y}^o)&=\log\int p(\mathbf{X})p(\mathbf{W})p(\mathbf{F}|\mathbf{X},\mathbf{U})p(\mathbf{Y}^o|\mathbf{F})\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{F}\mathrm{d}\mathbf{W}\\&\geq-\mathrm{KL}(q(\mathbf{X})\|p(\mathbf{X}))-\mathrm{KL}(q(\mathbf{W})\|p(\mathbf{W}))\\&\quad+\sum_n\sum_{d\in\mathcal{O}_n}\int q\left(\mathbf{x}_n\right)q\left(\mathbf{W}_d\right)p\left(\mathbf{f}_{nd}|\mathbf{x}_n,\mathbf{W}_d\right)\\&\quad\times\log p\left(\mathbf{y}_{nd}^o|\mathbf{f}_{nd}\right)\mathrm{d}\mathbf{x}_n\,\mathrm{d}\mathbf{f}_{nd}\,\mathrm{d}\mathbf{W}_d,\end{aligned} \tag{20}$$

where for the categorical variables, we write

$$p\left(\mathbf{f}_{n,d}|\mathbf{x}_n,\mathbf{W}_d\right)=\prod_{k=1}^{K}p\left(f_{n,d,k}|\mathbf{w}_{dk}\right), \tag{21}$$

and for the numerical variables,

$$p\left(\mathbf{f}_{n,d}|\mathbf{x}_n,\mathbf{W}_d\right)=p\left(f_{n,d}|\mathbf{w}_d\right). \tag{22}$$

The KL divergence terms in (20) penalize any deviation of the posterior from the prior of $\mathbf{X}$ and $\mathbf{W}$, respectively. Importantly, the KL divergence can be computed in a closed form under certain conditions [14]. The last term of the ELBO contributes to the reconstruction term of the observed data $\mathbf{Y}^o$. In order to compute the KL divergences, we consider Gaussian distributions

$$q(\mathbf{W})=\prod_{d=1}^{D_c}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{w}_{dk}|\mu_{dk},\mathbf{K}_d\right)\prod_{d'=D_c+1}^{D_c+D_q}\mathcal{N}\left(\mathbf{w}_{d'}|\mu_{d'},\mathbf{K}_{d'}\right), \tag{23}$$

$$q(\mathbf{X})=\prod_{n=1}^{N}\prod_{q=1}^{Q}\mathcal{N}\left(x_{nq}|m_{nq},\sigma_{nq}^2\right). \tag{24}$$

The parameters we need to optimize include μ_{dk}, $\mathbf{K}_d$, $\mu_{d'}$, $\mathbf{K}_{d'}$, m_{nq}, and σ_{nq}^2. In order to obtain a Monte Carlo estimate of the gradients with low variance, a useful trick introduced in [11] is to transform the random variables that need to be sampled so that their randomness does not depend on the parameters with which the gradients are computed.

For transforming $\mathbf{X}$, $\mathbf{w}_{d,k}$ ($\mathbf{w}_d$), and $\mathbf{f}_{n,d}$, we write

$$x_{ni}=m_{nq}+s_{nq}\varepsilon_{nq}^{(x)},\quad\varepsilon_{nq}^{(x)}\sim\mathcal{N}(0,1), \tag{25}$$

$$\mathbf{w}_{d,k}=\boldsymbol{\mu}_{d,k}+\mathbf{L}_d\varepsilon_{dk}^{(w)},\quad\varepsilon_{dk}^{(w)}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}_J\right) \tag{26}$$

where $\mathbf{L}_d\mathbf{L}_d^T=\mathbf{K}_d$ is the Cholesky decomposition of $\mathbf{K}_d$.

The problem of optimizing the ELBO becomes complex due to the presence of heterogeneous data. This leads to a problem with many local optima. These local optima capture the correlations between a subset of attributes while treating the rest as independent. However, the global optimum captures all the dependency among the attributes [8].

5. EXPERIMENTS AND RESULTS

In this section, we first evaluate the performance of our model in imputing the missing data with mixed numerical and categorical data, and we compare it to other imputation methods in the literature. We also study the classification task, where we evaluate the classification accuracy due to performing the imputation of the missing data in supervised scenarios. Note that our method does not require missing data imputation before the downstream task. Instead, the lower representation is obtained by integrating the missing data. We compare it with mean/mode imputation and one-hot encoding of categorical variables that are imputed using iterative imputation.

As a metric for performance comparison of the above models, we use the average imputation error computed as

$$AvgErr=1/D\sum_d\mathrm{err}(d), \tag{27}$$

where we use error metrics that are defined for the specific type of variables. More specifically, we adopt the accuracy error for categorical variables and the normalized root mean square error for numerical variables, defined by,

$$\text{err}(d) = \frac{1}{n}\sum_{n} I\left(y_{nd} \neq \hat{y}_{nd}\right), \quad d = 1 : D_c, \tag{28}$$

$$\text{err}(d) = \frac{\sqrt{1/n\sum_n \left(y_{nd} - \hat{y}_{nd}\right)^2}}{\max\left(\mathbf{y}_d\right) - \min\left(\mathbf{y}_d\right)}, \quad d = D_c + 1 : D_c + D_q. \tag{29}$$

5.1. Unsupervised (Missing Imputation)

We used the employee attrition dataset from IBM Watson Analysis. The data include 1470 samples that contain 33 features, including employees' age, gender, salary, job role, satisfaction, and their relationship to attrition. It has $D_c = 11$ categorical and $D_q = 22$ numerical variables. Table 1 summarizes the average imputation error as we vary the fraction of missing data. We compared against iterative imputation on one-hot encoded data and heterogeneous multi-output GP (HetMOGP). We observe that since mean imputation assumes all the features to be independent, any missing data imputation method that are based on statistical dependencies in the data will perform better than the mean imputation. The results show the ability of our method (we refer to it in the table as Heterogeneous Incomplete GP (HI-GP)) to exploit underlying correlations among the set of heterogeneous attributes.

Table 1. Average imputation error for different missing percentages

	10%	30%	50%
Mean Imputation	0.16 ± 0.02	0.16 ±0.02	0.20 ± 0.05
HetMOGP	0.15±0.03	0.15 ± 0.00	0.18 ± 0.02
One-hot/Iterative	0.15 ±0.01	0.15 ± 0.00	0.18 ± 0.01
HI-GP	0.13 ± 0.00	0.14 ± 0.00	0.17 ± 0.00

5.2. Supervised (Binary Classification)

Although the introduced generative model is fully unsupervised, we evaluate its performance in a binary classification problem. In this experiment, we again used the employee dataset. The idea behind this experiment is to treat the classes as missing values and try to predict/impute the class values. For example, if we use 60% of the data for training and 40% for testing, this means that we remove 40% of the labels in the target attribute and treat them as missing values. We compared the results against GP classification when the data are first imputed (IterGPC). The results for different percentages of missing values are summarized in Table 1. We observe that HI-GP provides better performance than that of Iter-GPC. They show that fully generative models like HI-GP are preferable over a supervised model with imputed data.

6. CONCLUSION

In this paper, we proposed a generative model for mixed categorical and numerical data using GPs. We note that this mixed-type problem combined with the presence of missing data has various challenges. In HI-GP, we derive a lower bound on the data marginal likelihood.

Table 2. Classification accuracy for different missing percentages

Missing %	Model	CA	F1	Precision	Recall
0%	HI-GP	0.711	0.710	0.713	0.711
	Iter-GPC	0.684	0.684	0.684	0.684
10%	HI-GP	0.705	0.704	0.706	0.705
	Iter-GPC	0.665	0.664	0.665	0.665
30%	HI-GP	0.650	0.649	0.650	0.650
	Iter-GPC	0.593	0.577	0.609	0.593

The ELBO depends only on observed data. Also, we study imputation in dealing with missing values. Our experiments showed that our proposed HI-GP outperformed its competitors on the missing data imputation task. Similarly, HI-GP provided better classification accuracy with supervised methods when compared to methods that cannot handle missing values in data, and therefore, require imputation of the missing inputs in the preprocessing stage. Future work includes the extension to more complex architectures of the generative model such as deep GPs or Bayesian neural networks.

7. REFERENCES

[1] Heidi Preis, Petar M Djurić, Marzieh Ajirak, Tong Chen, Vibha Mane, David J Garry, Cassandra Heiselman, Joseph Chappelle, and Marci Lobel, "Applying machine learning methods to psychosocial screening data to improve identification of prenatal depression: Implications for clinical practice and research," *Archives of Women's Mental Health*, pp. 1–9, 2022.

[2] Carl Edward Rasmussen and Christopher KI Williams, *Gaussian Processes for Machine Learning*, vol. 2, MIT press Cambridge, MA, 2006.

[3] Carl Edward Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[4] Neil D Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Advances in Neural Information Processing Systems*, 2004, pp. 329–336.

[5] Michalis Titsias and Neil D Lawrence, "Bayesian Gaussian process latent variable model," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 844–851.

[6] Marzieh Ajirak, Cassandra Heiselman, Anna Fuchs, Mia Heiligenstein, Kimberly Herrera, Diana Garretto, and Petar M Djurić, "Bayesian nonparametric dimensionality reduction of categorical data for predicting severity of covid-19 in pregnant women," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[7] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[8] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, vol. 107, pp. 107501, 2020.

[9] Vidhi Lalchand, Aditya Ravuri, and Neil D Lawrence, "Generalised Gaussian process latent variable models (GPLVM)

with stochastic variational inference," *arXiv preprint arXiv:2202.12979*, 2022.

[10] Pablo Moreno-Muñoz, Antonio Artés, and Mauricio Alvarez, "Heterogeneous multi-output Gaussian process prediction," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[11] Yarin Gal, Yutian Chen, and Zoubin Ghahramani, "Latent Gaussian processes for distribution estimation of multivariate categorical data," in *International Conference on Machine Learning*, 2015, pp. 645–654.

[12] Siddharth Ramchandran, Miika Koskinen, and Harri Lähdesmäki, "Latent Gaussian process with composite likelihoods and numerical quadrature," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3718–3726.

[13] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[14] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.