Online Learning in Variable Feature Spaces with Mixed Data

Yi He *, Jiaxian Dong †, Bo-Jian Hou ‡, Yu Wang †, Fei Wang ‡
* Department of Computer Science, Old Dominion University
† Institute of Artificial Intelligence and Blockchain, Guangzhou University
‡ Department of Population Health Sciences, Cornell University
*yihe@cs.odu.edu, †{jiaxiandong,yuwang}@gzhu.edu.cn ‡{boh4001, few2001}@med.cornell.edu

***Abstract*—This paper explores a new online learning problem where the data streams are generated from an over-time varying feature space, in which the random variables are of mixed data types including Boolean, ordinal, and continuous. The crux of this setting lies in how to establish the relationship among features, such that the learner can enjoy 1) reconstructed information of the missed-out old features and 2) a jump-start of learning new features with educated weight initialization. Unfortunately, existing methods mainly assume a linear mapping relationship among features or that the multivariate joint distribution could be modeled as a Gaussian, limiting their applicability to the mixed data streams. To fill the gap, we in this paper propose to model the complex joint distribution underlying mixed data with Gaussian copula, where the observed features with arbitrary marginals are mapped onto a latent normal space. The feature correlation is approximated in the latent space through an online EM process. Two base learners trained on the observed and latent features are ensembled to expedite convergence, thereby minimizing prediction risk in an online setting. Theoretical and empirical studies substantiate the effectiveness of our proposed approach. Code is released in https://github.com/xiexvying/OVFM.**

I. Introduction

The advent of Big Data has triggered a flurry of online machine learning algorithms that can mine hidden patterns from data streams. Whereas the initial focus of these algorithms was to deal with an ever increasing instance space – new data points arrive over-time from which a model is trained on-the-fly, recent advances have extended this traditional online learning paradigm to a novel setting of *doubly-streaming* data mining, where the data streams do not only increase their volume by having new instances, but also increase their *dimension* in terms of new features appearing. To learn such data, pioneer studies [1]–[4] prescribed that, once the new features appear, they shall persist in all later data instances, leading to a monotonically increasing feature space. Subsequent studies relaxed this setting by allowing the pre-existing features to be missed out afterwards with a batch-by-batch regularity [5]–[9] or in an arbitrary fashion [10]–[14], leading to a feature space that flexibly varies along the time horizon; Such relaxation generalizes the novel setting to a wider range of applications including sensor networks [15], novelty detection [16], cybersecurity [17], to name a few. This general learning paradigm has been termed as *online learning in variable feature spaces* (OLVFS) [14].

Despite effective, these prior arts mainly posit that all features, including the pre-existing and newly arriving ones, must come from the same data family. In other words, all data instances can contain digits of one data type only. Unfortunately, due to the messy and heterogenous nature of the real-world data, such an assumption is way too restrictive. Take, as a tangible example, the clinical data streams where the instances collected by various medical service providers can include data being Boolean (*e.g.*, skinny vs obese), ordinal (*e.g.*, 0–IV cancer stages), and continuous (*e.g.*, in-vivo insulin levels) and thus are of mixed types [18]. Enforcing the existing OLVFS algorithms to learn from such *mixed data streams* does not work well and shall be limited in two aspects as follows.

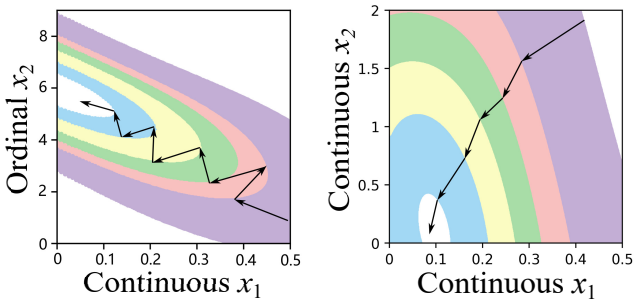


Fig. 1: Illustration of the optimization process over two 2D contours formed by 1) *left*: continuous vs ordinal variables and 2) *right*: two continuous variables. The ordinal variable introduces high feature variation thus garbles the gradients.

First, the crux of existing OLVFS methods lies in establishing the relationship among features, so that the knowledge learned from the old features could be leveraged to initiate a jump-start of learning new features. Yet, the joint distribution of mixed data can be complex and difficult to be delineated. For example, Gaussians which have been commonly used in the literature [5], [12], [14] fail to model the joint distribution between ordinal (discrete) and continuous variables. A plausible solution is to treat the ordinal data as generated from coarse binning of continuous data [19]. Alas, this solution does not work well either because 1) it sacrifices invaluable information of the orders; 2) it can afford very limited number of bins/categories only; and 3) it ignores the imbalanced data distribution and discriminative power across various bins[1].

[1] For example, the difference among cancer stages 0–III are trivial but substantially grow in stages IV and V; Therefore, the distance between stage 0 to I shall differ from that between III to V.

Second, mixed data streams result in high feature variation, leading to slow convergence rate and thereby in online classification incurring large *regret*. Figure 1 illustrates how the stochastic gradients are garbled by two uneven features being ordinal and continuous in the 2D loss contours, from which we observe that optimizing over the mixed data takes more steps to converge due to the aggressive updates along the direction of the ordinal variable. In contrast, continuous variables allow to tune the updating directions in a finer level of granularity hence yield faster convergence rate. Note, feature normalization techniques cannot generalize to doubly-streaming data as they prescribed either a fixed instance space (*i.e.*, offline pre-normalization) [20] or a fixed feature space [21].

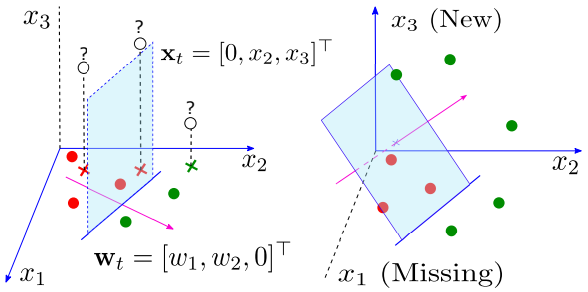
To overcome the two challenges, we in this paper propose a novel algorithm, termed *Online learning in Variable Feature spaces with Mixed data* (OVFM), which leverages Gaussian copula [22]–[24] to model the feature correlation in a *latent space*, encoded by continuous variables, without making assumption on the original distribution or data family of any feature. Specifically, the observed features (both ordinal and continuous) are deemed as generated from drawing the marginals of a latent normal vector after a coordinate-wise monotonic transformation. The ordinal features are thus associated with continuous variables via thresholding [25], where each ordinal level corresponds to an interval of values of the latent normal variable. An online learner trained on this latent space thus enjoys an improved performance over that trained on the observed features only, as it benefits from 1) being provided extra information rendered from reconstructing the missing features or initializing the learning weights for new features with educated guess and 2) eliminating aggressive updates encouraged by ordinal variables.

Specific contributions of this paper are as follows:

1) This is the first work to explore the problem of online learning in a feature space that varies its dimension over-time and includes mixed data types. The technical challenge of this new problem has been reasoned from a perspective of hyperplane optimization (in Section II).
2) A novel OVFM algorithm is proposed to tackle the new online learning problem through using Gaussian copula to establish the relationship between old and new features in a continuous latent space (in Section III).
3) A theoretical study in Section IV shows that i) the empirical estimators that infer the missing/new features from the old features are unbiased (Lemma 1 and Lemma 2) and ii) our approach provably enjoys a tighter regret bound than an online algorithm that trains learner on the observed features only (Theorem 1).
4) Extensive experiments are carried out over 14 datasets and the results demonstrate the viability, effectiveness, and superiority of our proposal (in Section V).

II. The Learning Problem

Let $\{(\mathbf{x}_t, y_t) \mid t = 1, 2, \ldots, T\}$ denote an input sequence, where $\mathbf{x}_t = [x_1, x_2, \ldots, x_{d_t}]^\top \in \mathbb{R}^{d_t}$ is a d_t-dimensional vector observed at the t-th round, accompanied with a label $y_t \in \{-1, +1\}$. In a variable feature space, d_i does not necessarily equate to d_j for any two time steps i and j. In the setting of mixed data, we let $\mathbf{x}_t := (\mathbf{x}_C, \mathbf{x}_D)$, where the subscripts C and D denote the continuous and discrete (*i.e.*, Boolean or ordinal) variables, respectively.



(a) Uneducated Guess (b) True Hyperplane

Fig. 2: Prediction errors incurred by a feature space changed at two consecutive rounds. (a) A learner that can correctly separate the green and red dots in the (x_1, x_2)-space now observes three new dots from the (x_2, x_3)-space, two of which are classified as red; (b) True hyperplane telling correct prediction suggests that all the three new dots should be classified as green.

At the t-th round, the learner ϕ_t observes $\mathbf{x}_t$ and makes prediction. An instantaneous loss indicating the discrepancy between the prediction and the true label is suffered, and the learner updates to ϕ_{t+1} based on the loss. Our goal is to find a series of functions $\phi_1, \ldots, \phi_T$ that predicts the data sequence accurately by means of empirical risk minimization (ERM) [26], defined as $\min_{\phi_1,\ldots,\phi_T} \frac{1}{T}\sum_{t=1}^{T} \ell\big(y_t, \phi_t(\mathbf{x}_t)\big)$, where $\ell(\cdot,\cdot)$ denotes a loss metric and often is prescribed as convex for simplicity such as square loss or logistic loss.

Mixed-Data Challenges and Our Thoughts

Two challenges impede the existing online learning algorithms to work well in our learning problem, described as follows.

Challenge I – *Modeling Feature-wise Relationship with Mixed Data:* Without loss of generality, we consider that the feature space can vary arbitrarily, where at any round new features appear and pre-existing features missed out without following any regularity. This setting leads to a highly dynamic learning environment that renders the learner's incapability of making accurate predictions. Take, as a simple example shown in Figure 2, that the mis-classification is incurred by the uneducated guess, where the learner trained in the (x_1, x_2)-space is enforced to predict the (x_2, x_3)-points. The fact that the learned information of x_1 cannot be used and there exists no prior knowledge of the new feature x_3 leads to a prediction of $\phi_t(\mathbf{x}_t) = \text{sign}(\mathbf{w}_t^\top \mathbf{x}_t) = \text{sign}(w_1 \cdot 0 + w_2 \cdot x_2 + 0 \cdot x_3)$, where only the shared feature x_2 is exploited, ending up with substantial prediction errors.

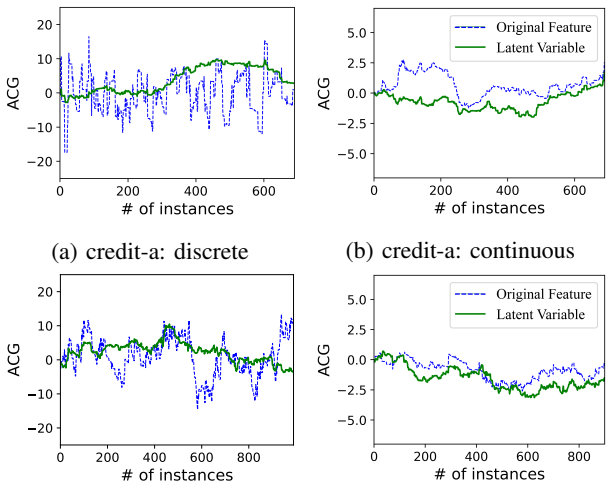
To aid the issue, existing OLVFS methods such as [2], [12], [14] proposed to establish the relationship between the observed features and the features being either missing or new. Such feature-wise relationship lend these methods to enjoy an educated approximation of the true hyperplane via 1) reconstructing the missing feature x_1 and 2) initializing the weight of new feature x_3 strategically.

Unfortunately, none of these methods can work well in a mixed-data setting. Let $\mathcal{U}_t \subseteq \mathbb{R}^{d_1} \cup \mathbb{R}^{d_2} \cup \ldots \cup \mathbb{R}^{d_t}$ denote a *universal feature space* that includes all appeared features up to the t-th round. Establishing the relationship among features is equated to learning a mapping $\psi : \mathbb{R}^{d_t} \mapsto \mathcal{U}_t$. We let $\psi(\mathbf{x}_t) := (\mathbf{x}_O, \mathbf{x}_M) = (\mathbf{x}_C, \mathbf{x}_D, \mathbf{x}_M)$, where $\mathbf{x}_O$ and $\mathbf{x}_M$ denote the observed and missing features, respectively, and $|\mathbf{x}_M| = |\mathcal{U}_t \setminus \mathbb{R}^{d_t}|$. Prior OLVFS methods all assume the conditional probability of $\mathbf{x}_M$ given $\mathbf{x}_O$ can be modeled by a continuous Gaussian and hence cannot deal with the discrete variables. A plausible solution is to treat ordinal data as continuous when establishing the mapping ψ and then apply cutoffs to bin the reconstructed missing variables into discrete categories. However, this solution can handle very limited number of categories and entails extensive expert knowledge to craft sensible cutoff thresholds. Also, this solution overlooks the disparateness and the ordering information existing across categories, thereby incurring substantial reconstruction error.

Challenge II – *Gradient-based Optimization over Discrete Variables:* To preserve the "online property" when minimizing the empirical risk, existing OLVFS algorithms all rely on the gradient information to perform stochastic updates. However, carrying out gradient-based optimization methods over ordinal data is non-trivial due to their discreteness. Intuitively, partial derivatives on discrete variables are in a coarser level of granularity than those on continuous variables and thus tend to encourage more radical updates.

A simple example reduced from the "credit-a" and "german" datasets in Section V is given in Figure 3 to rationalize this intuition. Specifically, online convex programming [27] is employed as the optimizer, and the variations of the averaged cumulative gradients (ACGs) [28] corresponding to the ordinal and continuous features along the time horizon are illustrated. We observe that the discrete variables in both datasets incur substantial turbulence of ACGs over time, which suggests aggressive updates during the online learning process. In contrast, the variation of ACGs corresponding to continuous features are smoother, finer-tuning the learner at small scale. As such, the stochastic updates tend to be garbled in the sense that, during learning, the updating directions that correspond to negative gradients of ordinal features dominate in the high-dimensional space, leading to large steps that walk in an oscillating fashion (as we can observe that the ACGs of ordinal features vibrate around the zero values). When such garbled updates accumulate, the online learner converges slowly, thereby tending to make more classification errors.

Our Idea: To tackle the above two challenges, we desire a model that is capable of 1) delineating the joint distribution between continuous and discrete variables, so as to establish feature-wise relationship in mixed data streams, and 2) normalizing the oscillating gradients over discrete dimensions into a continuous domain, so as to encourage finer updates and faster convergence rate. We advocate that Gaussian copula (GC) [22]–[24] provides such a model that possesses the two nice properties at once. Namely, GC can model the complex multivariate distribution of mixed data in a latent space spanned by continuous normal variables. An online learner is trained directly on the latent space, enjoying two-fold advantages.



(a) credit-a: discrete (b) credit-a: continuous

(c) german: discrete (d) german: continuous

Fig. 3: Trends of averaged cumulative gradients (ACGs) *w.r.t.* the number of rounds in two datasets, credit-a and German. Blue dash-line and green line represent the ACGs of the original features (discrete or continuous) and the latent variables modeled by GC (continuous normal), respectively.

First, the latent representations of the missing features can be reconstructed from those of the observed features via the GC-modeled distribution, such that the learned information of the missing features could be exploited, leading to more accurate predictions. Second, the latent variables are continuous and hence allow gradient-based updates in a fine-level of granularity, eliminating the garbled gradients caused by optimizing over discreteness and thereby encouraging faster convergence rate in an online ERM process (as shown in Figure 3).

III. The Proposed Approach

Overview. In a nutshell, our approach can be framed in the objectives taking the following form,

$$\min_{\phi_1,\ldots,\phi_T} \frac{1}{T}\sum_{t=1}^{T} \ell\big(y_t, \phi_t(\mathbf{z}_t)\big) + \Omega(\phi_t), \tag{1}$$

$$\max_{\mathbf{f};\Sigma} \mathrm{P}_{\mathbf{x}_t \in B}\left[\mathbf{x}_t \mid \mathbf{z}_t; \mathbf{f}^{-1}, \Sigma\right], \tag{2}$$

$$\text{s.t.} \quad (\mathbf{x}_C, \mathbf{x}_D, \mathbf{x}_M) \overset{\text{i.i.d.}}{\sim} \mathrm{GC}(\mathbf{z}_t; \mathbf{f}, \Sigma). \tag{3}$$

In this section, we extrapolate the objectives in a sequence as follows. i) The constraint Eq. (3) presumes that the data sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are independently drawn from an unknown distribution, which yet can be modeled by a Gaussian copula (GC), with its details presented in Section III-A. ii) The likelihood maximization function Eq. (2) estimates the parameters of GC in a buffer B through online Expectation-Maximization (EM); A latent space that can represent each input $\mathbf{x}_t$ (having continuous and discrete variables) with continuous normal vector $\mathbf{z}_t$ is learned. We scrutinize this part in Section III-B. iii) An online learner is trained on the latent representations as indicated in Eq. (1); A regularization term $\Omega(\phi_t)$ is imposed on the learner to encourage a sparse

solution, so as to deal with an infinitely growing feature space. We close this section by introducing an ensemble strategy that boosts our learner with a provably better performance. Technical details of this part are given in Section III-C.

A. Mixed-Data Gaussian Copula

We model the joint distribution underlying the mixed data streams with Gaussian copula (GC) for two purposes. First, at the t-th round, given an instance $\mathbf{x}_t := (\mathbf{x}_C, \mathbf{x}_D)$ in which each feature drawn from an arbitrary distribution, GC can capture the dependency structure among features and meanwhile respects the specific marginal of each variable independently. Note, this property empowers GC to construct a multivariate distribution for mixed data (*i.e.*, discrete vs continuous) in our setting, which excels, *e.g.*, multivariate Gaussian in which every variable has to be continuous that strictly follows a Gaussian.

Second, up to the t-th round ($t > 1$), because of the varying nature of the feature space in doubly-streaming data, each input $\mathbf{x}_t$ carries a subset of the features observed so far (*i.e.*, the $\mathcal{U}_t$). A learner trained on $\mathbf{x}_t$ hence suffers from the loss of information associated with the missing features and tends to perform inferiorly. GC aids this issue with its data reconstruction mechanism. Specifically, GC maps the observations onto a latent space that contains sufficient statistics to estimate the missing entries. A GC is formally defined as follows.

Definition 1 (Gaussian Copula, GC)**.** *For any random vector $\mathbf{x} \in \mathbb{R}^d$ that follows the Gaussian copula* $\mathrm{GC}(\mathbf{z}, \mathbf{f}, \Sigma)$*, there exists an element-wise monotone function $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^d$ and a correlation matrix Σ such that $\mathbf{f}(\mathbf{z}) = \mathbf{x}$ for $\mathbf{z} \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$.*

As such, the monotone $\mathbf{f}$ establishes the mapping between the observed vector $\mathbf{x}_t$ and its latent representation $\mathbf{z}_t$, where Σ specifies the normal distribution of $\mathbf{z}_t$. In addition, a nice property of GC is that the correlation Σ is invariant to element-wise strictly monotone transformation [29]. This property allows GC to deal with discrete variables in $\mathbf{x}_t$ with a monotone cutoff operator taken on probability mass functions. Specifically, for any ordinal variable $x_i \in \mathbf{x}_D$ with range $|k|$ and mass function $\{p_l\}_{l=1}^k$, the corresponding mapping $\mathbf{f}$ is defined as:

$$f_i := \mathrm{cutoff}(z; \mathrm{S}) = 1 + \sum_{s_l \in \mathrm{S}} \mathbb{1}(z > s_l) \tag{4}$$

where $z \in \mathbb{R}$ is a continuous random variable with cumulative distribution function (CDF) F_z and $\mathrm{S} = \{s_l = F_z^{-1}(\sum_{t=1}^{l} p_t) : l \in |k-1|\}$. So by the invertibility of a monotone function, the latent representation of $\mathbf{x}_t$ could be calculated as $\mathbf{f}^{-1}(\mathbf{x}_t) = \big(\mathbf{f}^{-1}(\mathbf{x}_C), \mathrm{cutoff}^{-1}(\mathbf{x}_D)\big)$, where we distinguish the discrete mapping $\mathrm{cutoff}(\cdot)$ from the monotone function set $\mathbf{f}(\cdot)$ for the sake of clarification. In particular, for a continuous variable, its latent representation takes a specified real-value; for a discrete variable. its latent representation lies in the Cartesian product of an interval. Note, the dimension of $\mathbf{f}^{-1}(\mathbf{x}_t)$ equates to that of $\mathbf{x}_t$ but does not match to that of $\mathcal{U}_t$. Below, we present how to construct $\mathbf{z}_t = \psi(\mathbf{x}_t) \in \mathbb{R}^{|\mathcal{U}_t|}$ by establishing relationships between the observed $(\mathbf{x}_O, \mathbf{x}_C)$ and the missing $\mathbf{x}_M$.

Feature Reconstruction with GC. The key idea here is to infer $\mathbf{x}_M$, the mismatch between the input $\mathbf{x}_t$ and the universal space $\mathcal{U}_t$, by mapping the conditional mean vector of the corresponding $\mathbf{z}_M$ via the marginals of the observed $\mathbf{x}_O$. The reconstruction takes two approximation steps, namely, 1) taking the expectation of the observed latent representation $\mathbf{z}_O$ given the observation $\mathbf{x}_O$ and 2) taking the expectation of missing latent representation $\mathbf{z}_M$ given $\mathbf{z}_O$, defined as follows:

$$\begin{aligned} \hat{\mathbf{z}}_M &= \mathbb{E}\big[\mathbb{E}[\mathbf{z}_M \mid \mathbf{z}_O, \Sigma] \mid \mathbf{x}_O, \Sigma\big] \\ &= \Sigma_{M,O}\Sigma_{O,O}^{-1}\mathbb{E}\left[\mathbf{z}_O \mid \mathbf{x}_O, \Sigma\right], \end{aligned} \tag{5}$$

where $\Sigma_{M,O}$ and $\Sigma_{O,O}$ denote the sub-matrices of correlation Σ with rows and columns corresponding to the feature indices of $(\mathbf{x}_M, \mathbf{x}_O)$ and $(\mathbf{x}_O, \mathbf{x}_O)$, respectively. By reconstructing this empirical $\hat{\mathbf{z}}_M$ from the conditional expectation Eq. (5), the latent representation of $\mathbf{x}_t$ enjoys a full view with its dimension matching that of $\mathcal{U}_t$, namely, $\mathbf{z}_t = \psi(\mathbf{x}_t) = (\hat{\mathbf{z}}_C, \hat{\mathbf{z}}_D, \hat{\mathbf{z}}_M)$, where $\hat{\mathbf{z}}_C = \mathbf{f}^{-1}(\mathbf{x}_C)$ and $\hat{\mathbf{z}}_D = \mathrm{cutoff}^{-1}(\mathbf{x}_D)$. As such, we can further sampling from the copula $\mathrm{GC}(\mathbf{z}_t, \mathbf{f}, \Sigma)$ to have an approximation of the input, denoted by $\mathbf{x}_t^{\mathrm{rec}} = (\hat{\mathbf{x}}_O, \hat{\mathbf{x}}_M)$. The discrepancy between the observed $\mathbf{x}_t$ and its reconstructed version $\hat{\mathbf{x}}_O$ could indicate how well the function $\mathbf{f}$ and the correlation Σ have been estimated. This allows us to optimize these parameters by adapting the EM theory in a stochastic hence online fashion, which is extrapolated in the next section.

B. Online Expectation-Maximization for Parameter Estimation

To estimate the monotone function $\mathbf{f}$ and its inversion $\mathbf{f}^{-1}$, we adapt the common steps from [19] defining that $f_i^{-1} = \Phi^{-1} \circ F_i$, where Φ is a standard normal CDF and F_i corresponds to the true CDF of the i-th feature, yet is in general unavailable. To solve the issue, we estimate its empirical version $\hat{F}_i$ in a buffer B in which, at each round, an incoming instance is joined in and the oldest input is popped out. The estimator for continuous variables is defined as follows.

$$\hat{f}_i^{-1}(x_i) = \Phi^{-1}\big(H \cdot \hat{F}_i(x_i)\big), \tag{6}$$

where the scale $H = |B|/(|B|+1)$ guarantees a finite output. For discrete variables, we could define the cutoff S^i as a special case of Eq. (6) by replacing the probability mass p_l^i of the i-th feature with its sample mean, defined as:

$$\mathrm{S}^i = \left\{ \Phi^{-1}\left(\frac{\sum_{t=1}^{|B|} \mathbb{1}\left(\mathbf{x}_t[i] \le l\right)}{|B|+1} \right), l \in |k-1| \right\}, \tag{7}$$

where $\mathbf{x}_t[i]$ denotes the i-th (discrete) feature of the t-th input.

To estimate the correlation matrix Σ, introducing the buffer B is also beneficial. Indeed, in an offline setting, the correlation Σ enjoys a closed-form solution, namely, $\Sigma = \mathbf{X}\mathbf{Z}^{\dagger}(\mathbf{Z}^{\dagger})^{\top}\mathbf{X}^{\top}$, where $\mathbf{X}$ and $\mathbf{Z}$ represent the input data matrix and its corresponding latent representation, respectively [29]. In our doubly-streaming setting, however, no such merit can be harnessed, suggesting that an iterative approximation algorithm should be considered. As such, we propose to estimate the empirical $\hat{\Sigma}$ through running online Expectation-Maximization (EM) in an iterative manner using the buffer B.

Specifically, we strive to maximize the likelihood that the observed entries (denoted by $\mathbf{X}_O$) of the buffered matrix $\mathbf{X}_B \in \mathbb{R}^{|\mathcal{U}_t| \times |B|}$ can be accurately reconstructed by taking

the conditional expectation of Σ. To disambiguate the notation, we denote $\Sigma^{(t-1)}$ as the empirical correlation obtained in the precedent round and $\hat{\Sigma}$ as the objective to be approximated at the current round. The log-likelihood function is defined as:

$$Q\left(\hat{\Sigma}; \Sigma^{(t-1)}, \mathbf{X}_O\right) := \frac{1}{|B|} \sum_{t=1}^{|B|} \mathbb{E}\left[\mathcal{L}(\hat{\Sigma}; \mathbf{x}_t, \mathbf{z}_t) \mid \mathbf{z}_t, \Sigma^{(t-1)}\right]$$
$$= \text{const} - \frac{1}{2} \log\det(\hat{\Sigma}) - \frac{1}{2} \operatorname{Tr}\left(\hat{\Sigma}^{-1} G(\Sigma^{(t-1)}, \mathbf{x}_t)\right), \quad (8)$$

with $\Sigma^{(0)}$ initialized as an identity matrix. Two steps iterate in an alternative fashion to maximize Eq. (8) as follows.

E-step. We compute the empirical $\mathbf{z}_t$ as a conditional expectation given $\mathbf{x}_t$ and $\Sigma^{(t-1)}$ using Eq. (5), and so that to express the likelihood $Q(\hat{\Sigma}; \Sigma^{(t-1)}, \mathbf{X}_O)$ in terms of $\hat{\Sigma}$ by replacing $G(\Sigma^{(t-1)}, \mathbf{x}_t) = \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\top \mid \mathbf{x}_t, \Sigma^{(t-1)}]$ in Eq. (8).

M-step. We solve $\tilde{\Sigma} = \arg\max_{\Sigma} Q(\Sigma; \Sigma^{(t-1)}, \mathbf{X}_O)$, which guarantees to increase the likelihood by the EM theory (*cf.* Chapter 3 in [30]). Then we follow the idea of [31] to replace the correlation at the current round as a harmonic sum of the correlation obtained from the last round $\Sigma^{(t-1)}$ and $\tilde{\Sigma}$. This treatment can produce a sequence $\Sigma^{(1)}, \ldots, \Sigma^{(T)}$ with smooth updates. However, we note that this sequence represents a series of local maximizer of the likelihood that, albeit converges monotonically, is unconstrained. To fit the empirical maximizer into a normal covariance, we resort it to an approximation as:

$$\hat{\Sigma} = P_{\mathcal{E}}\left((1-\gamma_t)\,\Sigma^{t-1} + \gamma_t \tilde{\Sigma}\right), \quad (9)$$

with $\gamma_t \in (0, 1]$ being a decaying step size and $P_{\mathcal{E}}(\cdot)$ scales the positive diagonal of the empirical $\hat{\Sigma}$ to 1.

Algorithm 1: The OVFM Algorithm

Initialize : Classifiers $\mathbf{w}_O$ and $\mathbf{w}_Z$, correlation Σ, ensemble factor $\alpha = 0.5$, and cumulative risks $R_O^T = R_Z^T = 0$.
Parameters : Buffer B, sparsity c, and endpoint ε.

for $t = |B|, \ldots, T$ **do**
- Receive a mixed data instance $\mathbf{x}_t = (\mathbf{x}_C, \mathbf{x}_D)$;
- Join $\mathbf{x}_t$ in B and establish $\mathrm{GC}(\mathbf{f}, \Sigma)$;
- Estimate $\mathbf{f}$ for continuous $\mathbf{x}_C$ and discrete $\mathbf{x}_D$ with Eq. (6) and Eq. (7), respectively;
- **repeat**
  - /* Esitimate $\hat{\Sigma}$ with EM */
  - **for** $t = 1, \ldots, B$ **do**
    - *E-step:* Replace $G(\Sigma^{(t-1)}, \mathbf{x}_t)$ in Eq. (8) with $\mathbf{z}_t = (\mathbf{f}^{-1}(\mathbf{x}_C), \mathrm{cutoff}^{-1}(\mathbf{x}_D), \hat{\mathbf{z}}_M)$ calculated via Eqs. (4) and (5);
    - *M-step:* Estimate $\hat{\Sigma}$ using Eq. (9);
- **until** convergence or $\|\hat{\Sigma} - \Sigma^{(t-1)}\|_{\text{Forb}} \le \varepsilon$;
- /* Predict the oldest input in B */
- Pop vector $\mathbf{x}_{t-|B|+1}$ and reconstruct its latent $\mathbf{z}_{t-|B|+1} = \left(\mathbf{f}^{-1}(\mathbf{x}_C), \mathrm{cutoff}^{-1}(\mathbf{x}_D), \hat{\mathbf{z}}_M\right)$;
- Predict the label as $\mathrm{sign}(\hat{y}_{t-|B|+1})$ using Eq. (10) ;
- Reveal the true label $y_{t-|B|+1}$;
- Suffer risks and accumulate R_O^T and R_Z^T;
- Reweigh coefficient α using Eq. (11);
- Update classifiers $\mathbf{w}_O$ and $\mathbf{w}_Z$ using SGD;
- Sparsify $\mathbf{w}_O$ and $\mathbf{w}_Z$ using Eq. (12);

C. Online Ensemble Learning with Latent Space

Thus far, we have presented how to reconstruct the missing features via establishing a feature-wise relationship with GC and how to estimate parameters of the GC in an online fashion using a buffer. Given an input $\mathbf{x}_t$, its reconstructed version in the latent space is denoted as $\mathbf{x}_t^{\text{rec}} := [\hat{\mathbf{x}}_C, \hat{\mathbf{x}}_D, \hat{\mathbf{x}}_M]^\top \in \mathbb{R}^{|\mathcal{U}_t|}$, where $\hat{\mathbf{x}}_C = \mathbf{f}(\hat{\mathbf{z}}_C)$ and $\hat{\mathbf{x}}_D = \mathrm{cutoff}(\hat{\mathbf{z}}_D)$ denote the observed continuous and discrete variables being mapped back from the latent space, respectively, and $\hat{\mathbf{x}}_M = \mathbf{f}(\hat{\mathbf{z}}_M)$ is the missing variables reconstructed from the conditional expectation given other observed features. The online learner trained with the reconstructed feature vector can enjoy a complete information so as to make accurate predictions. Intuitively, in a linear classification regime, we could define the prediction at the t-th round as $\hat{y}_t = \mathrm{sign}\left(\mathbf{w}_t^\top \mathbf{x}_t^{\text{rec}}\right)$, where $\mathbf{w}_t \in \mathbb{R}^{|\mathcal{U}_t|}$ is the classifier trained on $\mathcal{U}_t$.

However, this straightforward learning method may not work well and is limited in two aspects. First, although additional information can be provided via reconstructing $\hat{\mathbf{x}}_M$, the precision of the reconstruction is decided by the function set $\mathbf{f}$ and the correlation Σ, of which the approximation could be negatively affected by an improperly chosen buffer size $|B|$ or by the limited number of seen instances at initial rounds. As a result, the prediction accuracy would be deteriorated if $\hat{\mathbf{x}}_M$ has not been precisely reconstructed. Second, the reconstructed $\hat{\mathbf{x}}_D$ again consists of discrete variables, over which the optimization process tends to converge slowly, incurring more prediction errors in an online setting.

To address the two issues, we propose to ensemble two base predictions – one defined on the observed feature space, thereby eliminating the errors introduced by inaccurate reconstruction, and the other defined on the continuous latent space, thereby enjoying a finer-granular and faster convergence rate during optimization. The ensemble prediction is defined as follows.

$$\hat{y}_t = \alpha \left\langle \mathbf{w}_O, \mathbf{x}_t \right\rangle + (1-\alpha) \left\langle \mathbf{w}_Z, \mathbf{z}_t \right\rangle, \quad (10)$$

where $\mathbf{w}_O \in \mathbb{R}^{d_t}$ and $\mathbf{w}_Z \in \mathbb{R}^{|\mathcal{U}_t|}$ are the classifiers corresponding to the input $\mathbf{x}_t$ and the latent $\mathbf{z}_t = (\hat{\mathbf{z}}_C, \hat{\mathbf{z}}_D, \hat{\mathbf{z}}_M)$. The intuition behind Eq. (10) is to let the ensemble coefficient α decide the impact of the observed features and the mapped latent normal vector in making decisions. Denoted by $R_O^T = \sum_{t=1}^T \ell(y_t, \langle \mathbf{w}_O, \mathbf{x}_t \rangle)$ and $R_Z^T = \sum_{t=1}^T \ell(y_t, \langle \mathbf{w}_Z, \mathbf{z}_t \rangle)$ are the cumulative risks suffered by making predictions on $\mathbf{x}_t$ and $\mathbf{z}_t$ over T rounds, respectively. We update the coefficient α at the round $T+1$ based on the risk exponentials [12], [26]:

$$\alpha = \frac{\exp(-\tau R_O^T)}{\exp(-\tau R_O^T) + \exp(-\tau R_Z^T)}, \quad (11)$$

where $\tau = 2\sqrt{2 \ln 2 / T}$ is a turned parameter. In implementation, the loss function is defined as the cross-entropy.

Learning Sparse Classifiers. With new features keep arriving ceaselessly, the size of $\mathcal{U}_t$ may soon grow into an unmanageable

size, where estimating and storing the function set $\hat{\mathbf{f}}$ and the correlation $\hat{\Sigma}$ and learning the classifier $\mathbf{w}_t$ may lead to both computational and memory overheads. To restrict the dimensional growth of $\mathcal{U}_t$, we dynamically drop less informative features, defined as those associating with small values of the weight coefficients [32]. To do this, the base classifiers are projected onto an ℓ_1-ball at each round, defined as:

$$\mathbf{w}_p \leftarrow \min\{1, c/\|\mathbf{w}_p\|_1\}\mathbf{w}_p, \quad p \in \{O, Z\}, \tag{12}$$

which encourages sparsity – most values of the weight vectors shall concentrate to the several largest entries. The positive parameter c determines how sparse the resultant classifiers are. Only a subset of γ features are retained after projection, while the other features with trivial values are dropped [1], [14], [33]. The main steps of our OVFM algorithm are summarized in Algorithm 1.

IV. Theoretical Analysis

In this section, we demonstrate the theoretical merits of our OVFM approach. The main results are two-fold. First, we show that the empirical estimation of the monotone function $\hat{\mathbf{f}}$ is tightly bounded to the true monotone function $\mathbf{f}$ over both continuous and discrete variables in Section IV-A. Second, we analyze the regret bound to rationalize the usefulness of the ensemble strategy in Section IV-B, and the result validates the asymptotic property of our OVFM approach. Due to space limitation, proof sketches are provided with details deferred to an electronic companion (available as supplementary material).

A. Tightness of Empirical Estimation

The rationality of Eqs. (6) and (7) lies in the argument that the empirical estimation of the monotone function $\hat{\mathbf{f}}$ is a good approximation of the true $\mathbf{f}$ that establishes a mapping between the observed variables and the latent normal in a Gaussian copula. This argument is indeed true with a static, large enough dataset [34]; Yet, in an online setting with a buffer B, it remains unknown that whether the estimated $\hat{\mathbf{f}}$ converges to the true $\mathbf{f}$ or not. More specifically, *how large is the gap between* $\hat{\mathbf{f}}$ *and* $\mathbf{f}$. This section serves to answer this question. The approximation bounds of continuous and discrete variables are given in Lemma 1 and Lemma 2, respectively, as follows.

Lemma 1. *Given a continuous random vector* $\mathbf{x} \in \mathbb{R}^d$ *with CDF* $F(x)$ *that follows the Gaussian copula* $\mathbf{x} \sim \mathrm{GC}(\mathbf{z}, \mathbf{f}, \Sigma)$, *in which each variable satisfies* $x_i \equiv \hat{f}_i(z_i)$, *where* $f_i = F_i^{-1} \circ \Phi$. *Let* $m = \min_{j\in|B|} x_i^{(j)}$ *and* $M = \max_{j\in|B|} x_i^{(j)}$ *denote the smallest and largest values of the* i*-th observed variable in the buffer* B, *respectively, the strictly monotone function* $\hat{f}_i$ *in Eq. (6) satisfies*

$$\mathrm{P}\left(\sup_{m\le x\le M}\left|\hat{f}_i^{-1}(x) - f_i^{-1}(x)\right| > \epsilon\right) \le 2e^{-c_1\epsilon^2|B|}, \tag{13}$$

with ϵ *taking an arbitrary value in* $(a_1|B|^{-1}, b_1)$ *and* a_1, b_1, c_1 *being constants associated with* $F(m)$ *and* $F(M)$.

Proof Sketch: The result is immediated by applying the Dvoretzky-Kiefer-Wolfowitz (DFW) inequality [35] which bounds the gap between an empirical CDF and the true CDF. Leveraging the monotonicity of $F(x)$, we have $|B|^{-1} < \epsilon < K_1 \equiv \min\{F(m)/4, (1+F(M))/4\}$, leading to define $r_c = \frac{|B|}{|B|+1}\mathcal{F}_{|B|}(x)$ so that $r_c \in [F(m)/2, (1+F(M))/2]$. It follows that $\sup_{x\in[m,M]}\left|\Phi^{-1}(r_c) - \Phi^{-1}(F(x))\right| < 2\epsilon \cdot \sup_{r\in[F(m)/2,(1+F(M))/2]}\left|\left(\Phi^{-1}(r)\right)'\right| = K_2 \equiv 1/\min\left\{\phi\Big(\Phi^{-1}\big(F(m)/2\big)\Big), \phi\Big(\Phi^{-1}\big((F(M)+1)/2\big)\Big)\right\}$, where $\left(\Phi^{-1}(r)\right)' = 1/\phi\left(\Phi^{-1}(r)\right)$ and ϕ and Φ here are the PDF and CDF of a standard normal, respectively. Adjusting the constants for $2K_2|B|^{-1} < \epsilon < 2K_1K_2$, along with the definition of $\hat{f}_i^{-1}$ in Eq. (6), complete the proof. □

Lemma 2. *Given a ordinal random variable* $x_i \in \mathbf{x}_C$ *in range* $|k|$ *with probability mass function* $\{p_l\}_{l=1}^k$, *associated with a latent normal variable* $z_i \in \mathbb{R}$ *that satisfies* $f_i(z) := \mathrm{cutoff}(z_i; \mathrm{S}) = x_i$. *The empirical cutoff in Eq. (7) satisfies*

$$\mathrm{P}\left(\|\hat{\mathrm{S}}_i - \mathrm{S}_i\|_1 > \epsilon\right) \le 2^{|k|}e^{-c_2|B|\epsilon^2/(|k|-1)^2}, \tag{14}$$

with ϵ *taking an arbitrary value in* $\big((|k|-1)a_2|B|^{-1}, (k-1)b_2\big)$ *and* $a_2, b_2, c_2 > 0$ *being constants associated with the mass function* $\{p_l\}_{l=1}^k$.

Proof Sketch: The proof proceeds in three steps. *First*, we define $s_l^* = \Phi^{-1}\big(\sum_{i=1}^B \mathbb{1}\left(x^i \le l\right)/|B|\big)$ for $l \in |k-1|$, it is verified that $s_0^* = -\infty$ and $s_k^* = +\infty$, and $\Delta_l^* = \Phi(s_l^*) - \Phi(s_{l-1}^*) = \sum_{i=1}^{|B|}\mathbb{1}(x^i = l)/|B|$. Note that the sequence $|B|\Delta_1^*, \ldots, |B|\Delta_{|k|}^*$ is multinomially distributed with parameters B and $p_1, \ldots, p_k$. We borrow the Bretagnolle-Huber-Carol inequality [36] to have that, for any $\varepsilon > 0$, $\sum_{l=1}^{|k|}\left|\Delta_l^* - p_l\right| < \epsilon$ with probability at least $1 - 2^{|k|}e^{-\frac{1}{2}|B|\epsilon^2}$. *Second*, for each $l \in |k|$, $|\Phi(s_l^*) - \Phi(s_l)| \le \sum_{t=1}^{|k|}|\Delta_t^* - p_t| < \epsilon$. Take $\epsilon > |B|^{-1}$, we arrive at $\Phi(s_l) - 2\epsilon < \Phi(s_l^*)\cdot\big(|B|/(|B|+1)\big) = \sum_{i=1}^{|B|}\mathbb{1}(x^i \le l)/(|B|+1) < \Phi(s_l) + 2\epsilon$. *Third*, when $l \in |k-1|$, we have $p_1 \le \Phi(s_l) \le \sum_{t=1}^{k-1} p_t$, and by letting $\epsilon < K_1 \equiv \min\{p_1/4, p_k/4\}$, we have $p_1/2 \le \Phi(s_l^*)\cdot\big(|B|/(|B|+1)\big) \le 1 - p_k/2$. Therefore, $\|\hat{S}_i - S_i\|_1 = \sum_{l=1}^{k-1}|\hat{s}_l - s_l| \le 2(k-1)\epsilon/K_2$, where $K_2 = 1/\min\left\{\phi\left(\Phi^{-1}\left(\frac{p_1}{2}\right)\right), \phi\left(\Phi^{-1}\left(1 - \frac{p_k}{2}\right)\right)\right\}$. Adjusting the constants yields $\mathrm{P}\left(\|\hat{S}_i - S_i\|_1 > \epsilon\right) \le 2\exp\left\{-\frac{1}{8K_2^2}\cdot\frac{|B|\epsilon^2}{(|k|-1)^2}\right\}$, which completes the proof. □

Lemma 1 suggests that the empirical $\hat{\mathbf{f}}$ converges to the true $\mathbf{f}$ in sup norm and the gap is bounded by the observed domain in the buffer B. Lemma 2 illustrates that the cutoff estimator $\hat{S}_i$ approximates S_i for each discrete variable. The larger the buffer size $|B|$, the tighter the empirical estimations approxmate the true mapping function being $\mathbf{f}$ for continuous and cutoff for ordinal features. In an online setting where new data points incoming ceaselessly, we could choose a large B to ensure an unbiased estimation of the mapping functions.

B. Performance Bound

We compare the prediction performance of OVFM with the two base learners being trained on the observed and the latent feature spaces independently. The ensemble strategy lends our OVFM algorithm a nice property, described as follows.

Theorem 1. *Over T rounds, we have*

$$\sum_{t=1}^{T} R_t \leq \min\{R_O^T, R_Z^T\} + \sqrt{2T\ln 2}, \tag{15}$$

where $\sum_{t=1}^{T} R_t$ denotes the cumulative risk suffered by making the ensemble prediction defined in Eq. (10).

Proof Sketch: The proof is completed with four observations. *First*, we define a quantitative $Q_T = \exp(-\tau R_O^T) + \exp(-\tau R_Z^T)$, and it is verified that $Q_1 = 2$ and $\ln(Q_T/Q_1) = -\tau \min\{R_O^T, R_Z^T\} - \ln 2$. *Second*, by expanding $R_p^T = R_p^{T-1} + r_p^T, \quad p \in \{O, Z\}$, where r_p^T is the instantaneous risk suffered by $\mathbf{w}_O$ or $\mathbf{w}_Z$ at the T-th round, we arrive at $\ln(Q_T/Q_{T-1}) = \ln\left[\alpha \exp(-\tau r_O^T) + (1-\alpha)\exp(-\tau r_Z^T)\right]$, with α defined in Eq. (11). *Third*, we leverage the convexity of loss function and adapt the Hoeffding Inequality (*cf.* Appendix A.1.1 in [26]) to deduce $\ln\left[\alpha \exp(-\tau r_O^T) + (1-\alpha)\exp(-\tau r_Z^T)\right] \leq -\tau R_T + \tau^2/8$. *Forth*, over T rounds we have $\ln(Q_T/Q_{T-1}) + \ldots + \ln(Q_2/Q_1) = \ln(Q_T/Q_1) \leq -\tau \sum_{t=1}^{T} R_t + T \cdot (\tau^2/8)$. Collecting the above four observations, we have $\sum_{t=1}^{T} R_t \leq \min\{R_O^T, R_Z^T\} + (\tau/8)T + \ln 2/\tau$, in which plugging $\tau = 2\sqrt{2\ln 2/T}$ completes the proof. □

So by Theorem 1, it follows that $\lim_{T\to\infty}(\sqrt{2T\ln 2}/T) = 0$, indicating that our OVFM is *asymptotically no-regret* compared to the two base learners. In fact, if the latent normal space provides more discriminant information so that $R_O^T - R_Z^T > \sqrt{2T\ln 2}$, we can verify that $\sum_{t=1}^{T} R_t < R_O^T$, which suggests that the online classifier being assisted with the latent features modeled by the GC can provably perform better than that trained on the raw observations.

V. Experiments

In this section, we deliver empirical evidences to substantiate that our OVFM algorithm is a viable and effective solution to the problem of online learning in variable feature spaces with mixed data and excels in two doubly-streaming settings, namely, *trapezoidal* data streams, where later inputs include increasingly more features, and *capricious* data streams, where new features appear and old features miss-out arbitrarily. Section V-A introduces the studied datasets. Section V-B elaborates the experiment setup. Results and findings are given in Section V-C.

A. Datasets

Our evaluations are benchmarked on 14 datasets, including one synthetic dataset and 14 selected from the UCI repository [37], with their statistics summarized in Table I. To validate the generalizability of our OVFM algorithm, we select the datasets from a diverse range of applications including economy, kinesiology, bioinformatics, and so on.

We generate a mixed dataset with explicitly known feature correlation, which can be used to valid whether the Gaussian copula is capable of accurately modeling joint distribution among various data types. We generate a covariance matrix Σ with its diagonal sums up to 1 and a zero mean vector. From this multivariate normal, we apply $P_{\mathcal{E}}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^j \left(\mathbf{x}^j\right)^\top\right)$ to generate 2000 samples in a dimensionality of 18 including 6 Boolean, 6 ordinal, and 6 continuous features. For continuous variables, their values are randomly sampled; For discrete variables being Boolean or ordinal, their values are sampled from a Cartesian product of two intervals binned from $(F_{max}(x) - F_{min}(x))/p$ with p being 2 or 5, respectively. A weight vector $\mathbf{w}$ is generated from a normal, such that the classification label for each sample $\mathbf{x}$ is synthesized as $y = \text{sign}(\mathbf{w}^\top \mathbf{x})$. The simulation of feature space dynamics are kept the same as that for UCI datasets.

TABLE I: Characteristics of the studied datasets.

Dataset	#Inst.	#Feat.	Dataset	#Inst.	#Feat.
ionosphere	351	34	german	1000	24
wdbc	569	30	synthetic	2000	18
australian	690	14	splice	3190	60
credit-a	690	15	kr-vs-kp	3196	36
wbc	699	9	HAPT	7352	561
diabetes	768	8	magic04	19,020	10
dna	949	180	a8a	22,696	123

B. Experiment Setup

Compared Methods. We take three online learning competitors to evaluate the effectiveness and generalizability of our OVFM approach in various settings.

- FOBOS [38] sets up a baseline that trains online learner on the observed features directly. To adapt it to a variable feature space, we pad zero values to the missing entries. Projected subgradient method is devised to encourage a sparse solution, so that the redundant features with (nearly) zero coefficients could be truncated.
- OLSF [2] was tailored to deal with a monotonically increasing feature space. Its key idea is to strategically initialize the weight coefficients for the new features in a passive-aggressive manner, where the new coefficients are reweighed from the old features if the incoming features convey fresh information that affects the decision boundary and remain unchanged otherwise.
- OCDS [12] was crafted to perform online learning in an arbitrarily varying feature space, where the joint distribution of all historical features is modeled by a multivariate Gaussian, through which the missing features at each round are recovered so as to offer the learner a complete feature information.

Evaluation Protocol. To perform a fair comparison, the experiments are benchmarked in two doubly-streaming settings. For the UCI datasets, to simulate *trapezoidal data streams*, we follow the setting as in [2] to divide the each entire dataset into 10 batches, where in the i-th batch only the first $i * 10\%$ features would be retained (*i.e.*, the first data batch will retain the first 10% features and so forth). To simulate *capricious data streams*, we follow the setting as in [12] to randomly remove at most 50% features in each incoming instance. Moreover, to ensure a mixed data setup, for the 5 UCI datasets (*i.e.*, credit-a, svmguide3, german, australian, diabetes) that are naturally with mixed data, we do not further preprocess; For the rest 9 UCI datasets that have continuous data only, we divide the

TABLE II: Results of cumulative error rate (CER $\pm$ standard deviation) on 14 datasets, the lower the better, where random shuffling has repeated 10 times for cross validation. The best results are bold. • indicates our approach has a statistically significant better performance than the counterparts (hypothesis supported by *paired t-tests* at 95% significance level).

	Trapezoidal Data Streams			Capricious Data Streams		
Dataset	FOBOS	OLSF	OVFM	FOBOS	OCDS	OVFM
ionosphere	**.203 ± .000**	.330 ± .001•	.232 ± .000	.356 ± .000•	.278 ± .003•	**.248 ± .001**
wdbc	.176 ± .001•	.222 ± .000•	**.129 ± .011**	.157 ± .000•	.169 ± .004•	**.082 ± .000**
australian	.339 ± .000•	.404 ± .000•	**.275 ± .000**	.279 ± .000•	.277 ± .003•	**.217 ± .000**
credit-a	.363 ± .000•	.392 ± .000•	**.300 ± .000**	.297 ± .000•	.262 ± .002•	**.240 ± .000**
wbc	.068 ± .000	.332 ± .000•	**.066 ± .000**	.131 ± .000•	.102 ± .000•	**.078 ± .000**
diabetes	.392 ± .001•	.432 ± .000•	**.260 ± .000**	.423 ± .000•	.379 ± .000•	**.315 ± .000**
dna	.268 ± .000	**.220 ± .000**	.262 ± .000	.239 ± .000•	.261 ± .006•	**.199 ± .000**
german	.278 ± .000	.372 ± .000•	**.267 ± .000**	.375 ± .000•	.369 ± .002•	**.319 ± .000**
synthetic	.347 ± .000•	.391 ± .000•	**.326 ± .000**	.360 ± .000•	.333 ± .001•	**.267 ± .000**
splice	**.362 ± .000**	.388 ± .000	.380 ± .000	.390 ± .000•	.354 ± .003•	**.326 ± .000**
kr-vs-kp	**.308 ± .000**	.317 ± .000	.442 ± .001	.316 ± .000	.359 ± .003•	**.313 ± .000**
HAPT	.322 ± .001•	.362 ± .002•	**.291 ± .001**	.241 ± .001•	.229 ± .002•	**.193 ± .001**
magic04	.305 ± .001•	.314 ± .003•	**.285 ± .004**	.251 ± .000	.272 ± .003•	**.238 ± .000**
a8a	.285 ± .001•	.367 ± .005•	**.262 ± .001**	.239 ± .002•	.220 ± .004•	**.188 ± .002**

entire feature space into 3 blocks and convert one block to Boolean type and one to the ordinal type with median and mean thresholding, respectively. Cumulative error rate (CER) which calculates the classification accuracy in an online fashion, namely, CER $= (1/t)\sum_{i\leq t}[\![y_i \neq \text{sign}(\hat{y}_i)]\!]$, is employed for performance evaluation, where $[\![\cdot]\!]$ takes the value of 1 if the argument is true and 0 otherwise.

C. Comparative Results

Table II and Figure 4 present the classification error rates and the CERs *w.r.t.* the number of instances, respectively. From the results, we aim to answer the following research questions.

Q1. *Does our approach excel among the state-of-the-arts?*

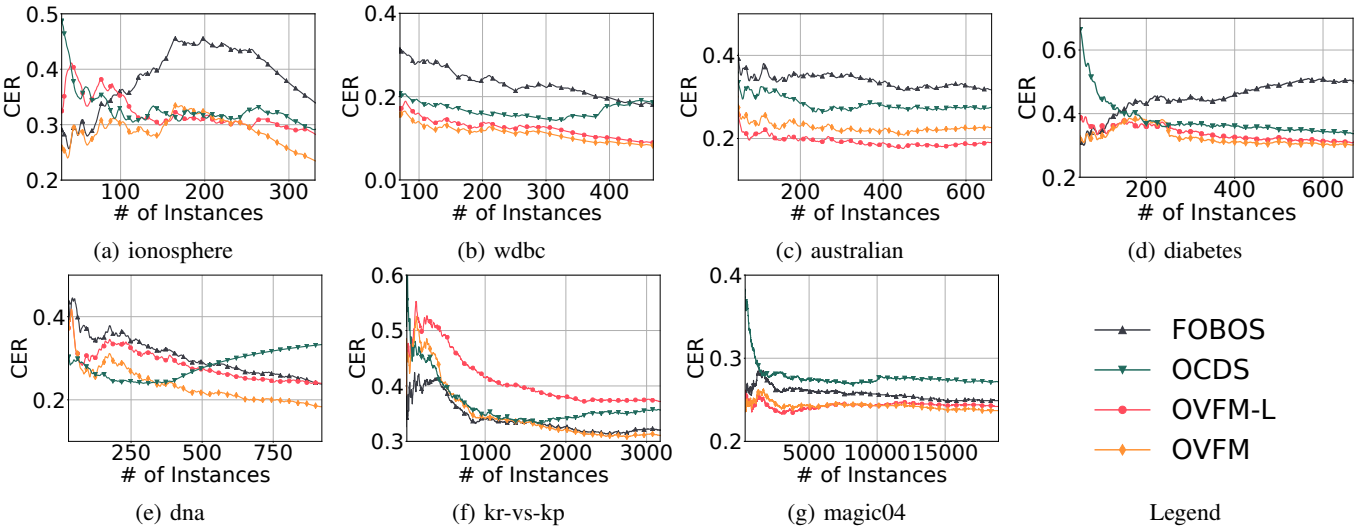
We make three observations from Table II to answer this question. *First*, our OVFM achieves the best performance with CERs of 26.9% and 23.0% on average in dealing with trapezoidal and capricious data streams, respectively. Supported by statistical evidence, OVFM excels in 45 out of 56 settings over the 14 studied datasets. *Second*, in the trapezoidal and capricious settings, OVFM outperforms the baseline FOBOS by ratios of 6.0% and 20.5%, respectively. This result suggests that the online learner in OVFM enjoys a latent space with full feature information and continuous data type over the observed features with mixed data, thereby performing robustly with various feature space dynamics. *Third*, although OLSF and OCDS have their respective mechanisms to deal with a variable feature space by establishing feature-wise relationship, our OVFM beats these two counterparts in their respective settings by ratios of 22.0% and 16.6% on average, respectively. The results substantiate the generalizability of using Gaussian copula to model the joint distribution over mixed-type variables in an online setting. This merit lets our approach enjoy robust performance across various feature space dynamics.

Q2. *How effectively can the Gaussian copula capture the relationship among features of mixed data types?*

The comparisons among FOBOS, OLSF, and OCDS amount to the answer. *First*, OLSF is inferior than FOBOS in 1 out of 14 datasets with their CER differing by ratio of 20.6%. Similarly, OCDS loses FOBOS in 4 settings with an on average 4.7% improved CER. The inferiority of OLSF and the neglectable improvement of OCDS are contrary to previous studies [2], [12] which have suggested a significant superiority of OLSF and OCDS over FOBOS in their respective scenarios. As the settings of the feature space dynamics are controlled as unchanged, we attribute such inconsistent results to the negative affect of mixed data. Specifically, OLSF and OCDS that prescribed the mapping relationships among features are linear and hence can be modeled trough a multivariate Gaussian do not fit well with mixed data streams, where the marginals of the continuous and discrete variables follow different distribution families and cannot be modeled with canonical frequentist models, *e.g.*, Gaussians. To further verify this, we in Figure 4 illustrate the trends of CER over time in the setting of capricious data streams, from which we observe that OCDS suffers from overfitting in Figures 4b, 4e, and 4f. The rising of CER after a certain time step indicate the incapability of linear models in capturing a nonlinear relationship between continuous and discrete variables. In contrast, both FOBOS, that simply assume a conditional independency of all variables given label, and our OVFM that employs Gaussian copula with the capability of modeling complex cross-type joint distribution do not suffer such performance degradation. These findings validate the effectiveness of Gaussian copula, which helped our OVFM approach to attain superior classification performance.

Q3. *Can ensemble learning boost classification accuracy?*

To answer this question, we compare OVFM with a variant named OVFM-L(atent), which trains online classifier and makes predictions on the latent space only. Indeed, the latent space is likely to yield fast and smooth convergence, because optimizing over continuous variables allows a finer model-tuning using gradients, as shown in Figure 4. However, in datasets where features show strong independency, forcibly modeling their correlation may lead to erroneously constructed latent variables and further negatively affect the classification accuracy, as shown in Figure 4f. As having the feature correlations in a prior is difficult in general and entails domain knowledge, ensemble learning is the strategy to lift this requirement so as



(a) ionosphere (b) wdbc (c) australian (d) diabetes

(e) dna (f) kr-vs-kp (g) magic04 Legend

Fig. 4: The trends of CER of 4 methods in the setting of capricious data streams on 7 datasets due to the page limits.

to suit datasets with various patterns. We observe that OVFM-L may excels in several, while our OVFM converge to lower error rate in almost all datasets, which validates the tightness of Theorem 1 and suggest the helpfulness of ensemble strategy in dealing with mixed data streams.

VI. Related Work

This work relates to online learning and copula modeling, so we review prior studies in each category and discuss the differences and relations. We note that, in concept-drift [39], [40] or non-stationary online learning [41] the changeables are the statistical properties of variables or the underlying decision function, respectively, as data streaming in, but the number of features carried by each input is fixed in a priori, which thus differ from our learning problem.

Online Learning in Variable Feature Space: Requiring an input sequence over wide time spans to be described by a fixed set of features is in general impractical. In response, the pioneer work [1]–[4] considered a doubly-streaming setting where new inputs carry consistently more features. Later studies extend this setting by allowing the pre-existing features to be missed out afterwards, either by following a batch-by-batch regularity [5]–[9] or at purely random [10]–[14]. This line of research shares the main idea of exploiting feature-wise relationship to anchor stationary information in a varied feature space. Unfortunately, existing methods all consider a continuous domain and thus do not scale up to the mixed data setting. Our OVFM filled the gap with Gaussian copula that maps and correlates arbitrary marginals across mixed variables and hence is more general.

Copula Modeling with Mixed Data: To explore a continuous latent space, most of methods assume that the data are of the same digital type and make a prior guess of the underlying distribution, as directly modeling the multivariate joint distribution for mixed marginals is difficult. In this regard, copulas lend us a tool due to its modeling power. Prior studies [34], [42], [43] proposed Gaussian copula by combining the cumulative distribution function (CDF) of each feature, and separating marginal distribution from multivariate distribution, which is suitable for modeling with different marginal distributions, harmonizing continuousness and discreteness. Subsequent works include Bayesian copula with factorized models [44], impute missing entries in a mixed data matrix by extending Rank-PC [45], and solving the time-varying complexities and high-dimensionality in estimating the covariance matrices [46]. However, most studies focus on modeling offline mixed data, with very few attempted online settings [19], [47]. None prior work has considered leveraging Gaussian copula to deal with an arbitrarily varying feature space so as to learn effective online learners; Our OVFM algorithm strived to fill the gap.

VII. Conclusions

In this paper, we explored a novel problem of online learning in variable feature space with mixed data. The challenge of this problem lies in the establishing of feature-wise relationship while the multivariate joint distribution of mixed data (*e.g.*, continuous vs discrete) is difficult to be modeled. We counter this challenge by proposing the OVFM approach that exploits Gaussian copula to model the mixed data with a latent space consisting of normal variables. A learner trained on this latent space enjoys 1) a complete feature information provided by the missing feature reconstruction and 2) a fast convergence rate since optimizing over the continuous latent variables renders gradients in a fine-level of granularity. A theoretical study demonstrated the performance advantages of our approach, and extensive empirical results further substantiated that.

Acknowledgement

We thank ICDM 2021 reviewers for their constructive feedback and Miss Xuying Xie and Mr. Shengda Zhuo, who are with the Institute of Artificial Intelligence and Blockchain, Guangzhou University, for their assistance in implementation of the proposed algorithm. The work done by Dr. Bo-Jian Hou and Dr. Fei Wang is supported by **... (Can please provide information here?)** The work done by Miss Jiaxian Dong and Dr. Yu Wang is supported by National Natural

Science Foundation of China (NSFC) under grant 61802080 and Guangzhou University Research Project under grants RQ2020085 and RD2020076. Any opinions, findings, and conclusions expressed in this publication are those of the authors and do not necessarily reflect the view of the funding agencies.

REFERENCES

[1] Q. Zhang, P. Zhang, G. Long, W. Ding, C. Zhang, and X. Wu, "Towards mining trapezoidal data streams," in *ICDM*. IEEE, 2015, pp. 1111–1116.

[2] ——, "Online learning from trapezoidal data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2709–2723, 2016.

[3] C. Hou, L.-L. Zeng, and D. Hu, "Safe classification with augmented features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2176–2192, 2018.

[4] E. Beyazit, M. Hosseini, A. Maida, and X. Wu, "Learning simplified decision boundaries from trapezoidal data streams," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 508–517.

[5] B.-J. Hou, L. Zhang, and Z.-H. Zhou, "Learning with feature evolvable streams," in *NeurIPS*, 2017, pp. 1417–1427.

[6] C. Hou and Z.-H. Zhou, "One-pass learning with incremental and decremental features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2776–2792, 2017.

[7] Z.-Y. Zhang, P. Zhao, Y. Jiang, and Z.-H. Zhou, "Learning with feature and distribution evolvable streams," in *ICML*. PMLR, 2020, pp. 11 317–11 327.

[8] B.-J. Hou, Y.-H. Yan, P. Zhao, and Z.-H. Zhou, "Storage fit learning with feature evolvable streams," in *AAAI*, 2021.

[9] B.-J. Hou, L. Zhang, and Z.-H. Zhou, "Prediction with unpredictable feature evolution," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

[10] E. Manzoor, H. Lamba, and L. Akoglu, "xstream: Outlier detection in feature-evolving data streams," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1963–1972.

[11] E. Beyazit, J. Alagurajah, and X. Wu, "Online learning from data streams with varying feature spaces," in *AAAI*, vol. 33, no. 01, 2019, pp. 3232–3239.

[12] Y. He, B. Wu, D. Wu, E. Beyazit, S. Chen, and X. Wu, "Online learning from capricious data streams: a generative approach," in *IJCAI*, 2019, pp. 2491–2497.

[13] ——, "Toward mining capricious data streams: A generative approach," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1228–1240, 2021.

[14] Y. He, X. Yuan, S. Chen, and X. Wu, "Online learning in variable feature spaces under incomplete supervision," in *AAAI*, vol. 35, no. 5, 2021, pp. 4106–4114.

[15] Y. Zhang, Y. Chen, H. Yu, X. Yang, R. Sun, and B. Zeng, "A feature adaptive learning method for high-density semg-based gesture recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–26, 2021.

[16] Z. Donyavi and S. Asadi, "Using decomposition-based multi-objective evolutionary algorithm as synthetic example optimization for self-labeling," *Swarm and Evolutionary Computation*, vol. 58, p. 100736, 2020.

[17] Y.-F. Li, Y. Gao, G. Ayoade, H. Tao, L. Khan, and B. Thuraisingham, "Multistream classification for cyber threat data with heterogeneous feature space," in *WWW*, 2019, pp. 2992–2998.

[18] H. Rolka, H. Burkom, G. F. Cooper, M. Kulldorff, D. Madigan, and W.-K. Wong, "Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs," *Statistics in Medicine*, vol. 26, no. 8, pp. 1834–1856, 2007.

[19] Y. Zhao and M. Udell, "Missing value imputation for mixed data via gaussian copula," in *KDD*, 2020, pp. 636–646.

[20] P. Ilmonen, H. Oja, and R. Serfling, "On invariant coordinate system (ics) functionals," *International Statistical Review*, vol. 80, no. 1, pp. 93–110, 2012.

[21] S. Ross, P. Mineiro, and J. Langford, "Normalized online learning," in *UAI*, 2013.

[22] J. Fan, H. Liu, Y. Ning, and H. Zou, "High dimensional semiparametric latent graphical model for mixed data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 2, pp. 405–421, 2017.

[23] P. D. Hoff *et al.*, "Extending the rank likelihood for semiparametric copula estimation," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 265–283, 2007.

[24] H. Liu, J. Lafferty, and L. Wasserman, "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research*, vol. 10, no. 10, 2009.

[25] H. Feng and Y. Ning, "High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference," in *AISTATS*. PMLR, 2019, pp. 654–663.

[26] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[27] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML*, 2003, pp. 928–936.

[28] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[29] G. Masarotto, C. Varin *et al.*, "Gaussian copula marginal regression," *Electronic Journal of Statistics*, vol. 6, pp. 1517–1549, 2012.

[30] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[31] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[32] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys*, vol. 50, no. 6, p. 94, 2017.

[33] J. Wang, P. Zhao, S. C. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 698–710, 2013.

[34] F. Durante and C. Sempi, "Copula theory: an introduction," in *Copula theory and its applications*. Springer, 2010, pp. 3–31.

[35] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

[36] A. W. Van Der Vaart and J. A. Wellner, "Weak convergence," in *Weak convergence and empirical processes*. Springer, 1996, pp. 16–28.

[37] D. Dua and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[38] J. C. Duchi and Y. Singer, "Efficient learning using forward-backward splitting." in *NeurIPS*, vol. 22, 2009, pp. 495–503.

[39] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[40] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[41] L. Zhang, "Online learning in changing environments." in *IJCAI*, 2020, pp. 5178–5182.

[42] M. Haugh, "An introduction to copulas," *IEOR E4602: quantitative risk management. Lecture notes. Columbia University*, 2016.

[43] D. Lopez-Paz, J. M. Hernández-Lobato, and G. Zoubin, "Gaussian process vine copulas for multivariate dependence," in *ICML*. PMLR, 2013, pp. 10–18.

[44] J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas, "Bayesian gaussian copula factor models for mixed data," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 656–665, 2013.

[45] R. Cui, P. Groot, and T. Heskes, "Robust estimation of gaussian copula causal structure from mixed data with missing values," in *ICDM*. IEEE, 2017, pp. 835–840.

[46] D. Salinas, M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus, "High-dimensional multivariate forecasting with low-rank gaussian copula processes," in *NeurIPS*, 2019, p. 6824.

[47] E. Landgrebe, M. Udell *et al.*, "Online mixed missing value imputation using gaussian copula," in *ICML – Workshop on the Art of Learning with Missing Values (Artemiss)*, 2020.