STABLEMISS+: PREDICTION WITH INCOMPLETE DATA UNDER AGNOSTIC MASK DISTRIBUTION SHIFT

*Yichen Zhu, Bo Jiang**

Shanghai Jiao Tong University
{zyc_ieee, bjiang}@sjtu.edu.cn

ABSTRACT

Missing data is ubiquitous in real-world scenarios. Recently, increasing attention has been given to prediction using only incomplete features together with a mask indicating the missing pattern. In this paper, we consider prediction with incomplete feature in the presence of distribution shift. In particular, we focus on the case where the joint distribution of complete feature and label is invariant, but the mask distribution may shift agnostically between training and testing. StableMiss is state-of-the-art in this problem. It removes correlations among feature, those among mask and those between feature and mask to avoid learning the correlations that possibly change under mask distribution shift. However, the correlations among feature can be helpful to prediction, since they do not change under mask distribution shift, and the optimal predictor, namely conditional expectation of label given incomplete feature, depends on them. To address this issue, we preserve the correlations among feature and simultaneously remove those among mask and those between feature and mask. Extensive experiments show that our method outperforms the state-of-the-art methods, with 10% reduction in RMSE.

***Index Terms*—** Prediction, incomplete data, agnostic mask distribution shift

1. INTRODUCTION

Missing data is ubiquitous in real-world scenarios due to sensor malfunction, incomplete sensing coverage, etc. Recently, increasing attention has been given to prediction with only incomplete feature, which consists of observed feature values and a mask that indicates which features are observed. Although most existing methods [1, 2] assume identical training and testing distributions, the prediction may take place in the presence of distribution shift. As noted in [3], a typical scenario is where the joint distribution of complete feature and label is assumed invariant between training and testing, but the mask distribution can be different, resulting from different sensor deployment, data management, etc. E.g., the naturally incomplete urban traffic speed dataset [4] throughout one year has relatively stable speed distribution, as the traffic network is almost unchanged, but its missing rate drops from 37% to 23% during this year. Moreover, the mask distribution shift is agnostic, since testing distribution is usually unavailable during the training process in practice. In this paper, we study prediction with incomplete feature under such agnostic mask distribution shift.

Several methods [1, 2, 5, 6, 7, 8] for missing data can be used for prediction with incomplete feature. However, they assume identical training and testing distributions and can hardly generalize under distribution shift. Several other methods [9, 10, 11, 12, 13] focus on prediction under agnostic feature distribution shift. They are designed for complete data and are not applicable to incomplete data.

StableMiss [3] has achieved state-of-the-art performance in our problem. It observes that the optimal predictor, namely conditional expectation of label given incomplete feature, is invariant between training and testing and proposes a prediction framework to approximate the optimal predictor. In order to make the learned model independent of mask distribution, StableMiss adapts existing decorrelation technique to incomplete data to avoid learning the correlations among feature and mask that possibly change under mask distribution shift. In particular, it removes three kinds of correlations, correlations among feature (intra-feature correlations), correlations among mask (intra-mask correlations) and correlations between feature and mask (inter-correlations). However, the intra-feature correlations can be helpful to prediction, since they do not change with mask distribution shift and the optimal predictor, namely conditional expectation of label given incomplete feature, depends on them.

In this paper, we address the above issue by preserving the intra-feature correlations and simultaneously removing intra-mask correlations and inter-correlations. Similar to [3], we decorrelate by sample reweighting that learns a weighting function of incomplete feature. Specifically, we first show the existence of weighting function that simultaneously preserves and removes the corresponding correlations when the mask is independent of the missing feature values. When the mask depends on the observed feature values, we learn the weighting function by regularizing complete feature distribution to be invariant and minimizing both intra-mask correlations and inter-correlations. When the mask is independent of feature, removing the intra-mask correlations automatically preserves the intra-feature correlations, and we learn the weighting function by only minimizing the intra-mask correlations. Finally, following [3], we conduct weighted regression under the training distribution. Our improved version of StableMiss is named StableMiss+.

The contributions of this paper are summarized as follows.

- This paper proposes a novel method StableMiss+ for prediction with incomplete feature under agnostic mask distribution shift. StableMiss+ can preserve intra-feature correlations and simultaneously remove infra-mask correlations and inter-correlations.
- Extensive experiments on both synthetic and real-world datasets show that StableMiss+ outperforms the state-of-the-art methods under agnostic mask distribution shift.

2. NOTATION AND PROBLEM STATEMENT

2.1. Notation

Capital and lowercase letters, e.g., X and x, are used to denote random variable and realization, respectively. Subscripts are used to

*Bo Jiang is the corresponding author.

This work is supported in part by the National Natural Science Foundation of China under Grant No. 62072302.

index the entries of a vector, e.g., x_i is the i-th entry of $\mathbf{x}$. Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^d$ denote the feature and label, respectively. We consider the case where $\mathbf{x}$ is partially observed and $\mathbf{y}$ is fully observed during training. We use a binary mask $\mathbf{m} \in \{0,1\}^n$ to indicate which entries of $\mathbf{x}$ are observed: $m_i = 1$ if x_i is observed, and $m_i = 0$ if x_i is missing. The complementary mask $\overline{\mathbf{m}}$ is defined by $\overline{m}_i = 1 - m_i, \forall i$. With a slight abuse of notation, we regard $\mathbf{m}$ and $\overline{\mathbf{m}}$ as the index sets of the observed and missing entries, so that the observed and missing feature values are $\mathbf{x}_\mathbf{m} = \{x_i \mid i \in \mathbf{m}\}$ and $\mathbf{x}_{\overline{\mathbf{m}}} = \{x_i \mid i \in \overline{\mathbf{m}}\}$, respectively. We consider the case where mask $\mathbf{m}$ is known, since it is common to know which features are observed within incomplete feature. The incomplete feature is given by $(\mathbf{x}_\mathbf{m}, \mathbf{m})$. We consider the case where label generation process depends on the feature but not the mask, i.e., $p(\mathbf{y} \mid \mathbf{x}, \mathbf{m}) = p(\mathbf{y} \mid \mathbf{x})$.

Following [14], we model the generative process of incomplete feature as follows. A complete feature sample $\mathbf{x}$ is first drawn from the complete feature distribution $p(\mathbf{x})$. Given $\mathbf{x}$, a mask sample $\mathbf{m}$ is then drawn from the conditional mask distribution $p(\mathbf{m} \mid \mathbf{x})$. The incomplete feature $(\mathbf{x}_\mathbf{m}, \mathbf{m})$ follows the distribution

$$p(\mathbf{x}_\mathbf{m}, \mathbf{m}) = \int p(\mathbf{x}) p(\mathbf{m} \mid \mathbf{x}) d\mathbf{x}_{\overline{\mathbf{m}}}.$$

We focus on the cases of Missing Completely At Random (MCAR) and Missing At Random (MAR) [14]. Under MCAR, mask $\mathbf{M}$ is independent of the underlying complete feature $\mathbf{X}$, i.e., $p(\mathbf{m} \mid \mathbf{x}) = p(\mathbf{m}), \forall \mathbf{m}, \mathbf{x}$; under MAR, mask $\mathbf{M}$ only depends on the observed feature values $\mathbf{X}_\mathbf{M}$, i.e., $p(\mathbf{m} \mid \mathbf{x}) = p(\mathbf{m} \mid \mathbf{x}_\mathbf{m}), \forall \mathbf{m}, \mathbf{x}$.

2.2. Problem Statement

Given a training set $\mathcal{D} = \{(\mathbf{x}^{(i)}_{\mathbf{m}^{(i)}}, \mathbf{m}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, consisting of N samples from the training distribution $p^{tr}(\mathbf{x}, \mathbf{m}, \mathbf{y})$, the goal is to learn a prediction function $g(\mathbf{x}_\mathbf{m}, \mathbf{m})$ for agnostic testing distribution $p^{te}(\mathbf{x}, \mathbf{m}, \mathbf{y})$, where the input to g is only incomplete feature. Under commonly adopted L_2-norm metric, the optimal g is

$$g(\mathbf{x}_\mathbf{m}, \mathbf{m}) = \mathbb{E}_{\mathbf{Y} \sim p^{te}_{\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}}}[\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}].$$

Following [3], we make the following assumption on testing distribution. Note that the mask distribution shift still remains agnostic under this assumption.

Assumption 1. *The joint distribution of complete feature and label is invariant between training and testing:*

$$p^{te}(\mathbf{x}, \mathbf{y}) = p^{tr}(\mathbf{x}, \mathbf{y}).$$

3. BACKGROUND

In this section, we introduce StableMiss [3], which is highly related to our method. It observes that, under Assumption 1 and in MCAR or MAR, the optimal predictor, i.e., conditional expectation of label given incomplete feature, is invariant between training and testing:

$$\mathbb{E}_{\mathbf{Y} \sim p^{te}_{\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}}}[\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}] = \mathbb{E}_{\mathbf{Y} \sim p^{tr}_{\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}}}[\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}].$$

StableMiss achieves generalization by approximating this invariant conditional expectation, whose architecture is shown in Fig. 1.

The conditional expectation $\mathbb{E}[\mathbf{Y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}]$ is an aggregation of 2^n optimal predictors, one for each mask $\mathbf{m}$. StableMiss approximates them by learning a prediction function g, whose parameter $\boldsymbol{\phi}$ is a function of $\mathbf{m}$ and parameterized by learnable $\boldsymbol{\theta}$, as given by

$$g(\mathbf{x}_\mathbf{m}, \mathbf{m}) = g_{\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{m})}(\mathbf{x} \odot \mathbf{m}),$$

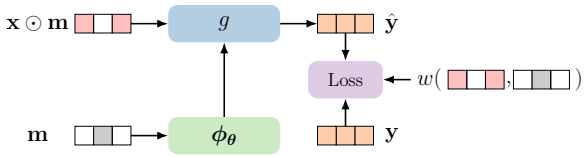

Fig. 1: Overall architecture of StableMiss [3].

where $\odot$ is element-wise multiplication.

To avoid learning the correlations among feature and mask that possibly change under mask distribution shift, StableMiss performs decorrelation by sample reweighting. Specifically, it learns a weighting function $w(\mathbf{x}_\mathbf{m}, \mathbf{m})$ of incomplete feature by minimizing the total correlations under the weighted training distribution $p_w(\mathbf{x}, \mathbf{m}, \mathbf{y}) = w(\mathbf{x}_\mathbf{m}, \mathbf{m}) p^{tr}(\mathbf{x}, \mathbf{m}, \mathbf{y})$. In the case of MCAR, it minimizes the intra-feature and intra-mask correlations:

$$\min_{\mathbf{w} \in \mathbb{R}^+} \sum \mathrm{cor}(X_k, X_l, \mathbf{w}) + \sum \mathrm{cor}(M_k, M_l, \mathbf{w}) + \gamma \mathrm{CV}(\mathbf{w}),$$

where the summation is taken over $1 \le k < l \le n$, and the coefficient of variation $\mathrm{CV}(\mathbf{w})$ of $\mathbf{w}$ is for regularization; in the case of MAR, it further minimizes the inter-correlations $\sum_{1 \le k, l \le n} \mathrm{cor}(X_k, M_l, w)$. The correlation $\mathrm{cor}(\cdot, \cdot, \mathbf{w})$ is measured by Random Fourier Feature [15]. Due to space limit, see [3] for details. Then it conducts weighted regression under the training distribution:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{p^{tr}}[w(\mathbf{X}_\mathbf{M}, \mathbf{M})(Y - g_{\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{M})}(\mathbf{X}_\mathbf{M}, \mathbf{M}))^2].$$

4. METHODOLOGY

4.1. Issue of Decorrelation

To make the model independent of mask distribution, StableMiss [3] aims to remove all the correlations among feature and mask, including infra-feature correlations. However, the infra-feature correlations can be helpful to prediction, whose reason is twofold. On the one hand, under Assumption 1, the intra-feature correlations do not change under mask distribution shift, and thus preserving them will not degrade the generalization performance as the intra-mask correlations and inter-correlations. On the other hand, the optimal predictor depends on the intra-feature correlations. The dependence can be seen by factorizing $p(\mathbf{y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m})$ as

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}) &= \int p(\mathbf{y} \mid \mathbf{x}_{\overline{\mathbf{m}}}, \mathbf{x}_\mathbf{m}, \mathbf{m}) p(\mathbf{x}_{\overline{\mathbf{m}}} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}) d\mathbf{x}_{\overline{\mathbf{m}}} \\ &= \int p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}_{\overline{\mathbf{m}}} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}) d\mathbf{x}_{\overline{\mathbf{m}}}. \end{aligned}$$

In the case of MCAR or MAR where $\mathbf{X}_{\overline{\mathbf{M}}}$ and $\mathbf{M}$ are conditionally independent given $\mathbf{X}_\mathbf{M}$, the second term $p(\mathbf{x}_{\overline{\mathbf{m}}} \mid \mathbf{x}_\mathbf{m}, \mathbf{m}) = p(\mathbf{x}_{\overline{\mathbf{m}}} \mid \mathbf{x}_\mathbf{m})$, which depends on the intra-feature correlations. Thus the optimal predictor depends on the infra-feature correlations. How we exploit the intra-feature correlations is detailed in the next section.

4.2. Improvement on Decorrelation

We exploit the intra-feature correlations by preserving them and simultaneously removing intra-mask correlations and inter-correlations. We first show in Theorem 1 below that theoretically this can be achieved by weighting the training distribution appropriately. Superscript tr of training distribution will be omitted for simplicity.

Table 1: Performance on Gaussian-Mix when trained under 50% missing level. The values for Optimal are RMSE, while the other values are the gap between the RMSE of the corresponding method and that of Optimal with the same experimental setup. **Bold** and underline represent the best and second best along each column, respectively. Superscript * indicates when training and testing missing levels are the same.

Method		Testing Missing Level									
		MCAR					MAR				
		10%	**30%**	**50%**	**70%**	**90%**	**10%**	**30%**	**50%**	**70%**	**90%**
Gap to Optimal	Partial VAE	1370.80	1022.81	872.67*	1058.23	1299.88	1541.52	1172.68	793.32*	1075.69	1112.42
	MIWAE	1141.75	977.36	825.64*	953.03	1019.08	1331.95	1047.14	743.63*	820.39	844.17
	P-BiGAN	1249.11	948.68	794.86*	942.71	1336.18	1382.71	1041.63	713.85*	845.39	979.45
	NeuMiss	607.77	480.56	312.75*	527.64	665.30	714.19	523.10	293.14*	601.74	679.99
	DWR	1290.18	931.87	873.53*	1072.58	1132.42	1485.42	1142.32	893.64*	969.56	963.52
	SRDO	1187.14	932.88	826.52*	1021.75	1217.66	1385.00	1080.93	843.85*	899.21	866.09
	StableNet	1061.32	833.19	725.82*	909.54	1190.70	1256.32	989.67	743.25*	807.50	788.56
	StableMiss	409.74	346.55	292.11*	331.50	403.29	431.16	309.09	282.75*	405.60	467.98
	StableMiss+	**378.92**	**320.46**	**290.85***	**309.34**	**369.57**	**408.90**	**285.39**	**277.86***	**381.07**	**454.95**
Optimal		861.06	1134.42	1301.95	1591.95	1795.19	904.87	1108.73	1324.00	1518.59	1705.54

Theorem 1. *Under Assumption 1 and in the case of MCAR or MAR, there exists a weight function $w(\mathbf{x_m}, \mathbf{m})$ such that, under the weighted training distribution $p_w(\mathbf{x}, \mathbf{m}) = w(\mathbf{x_m}, \mathbf{m})p(\mathbf{x}, \mathbf{m})$, the feature distribution is preserved, while intra-mask correlations and inter-correlations are removed, i.e.,*

$$p_w(\mathbf{x}) = p(\mathbf{x}),\ p_w(\mathbf{m}) = \prod_{i=1}^{n} p_w(m_i),\ p_w(\mathbf{m} \mid \mathbf{x}) = p_w(\mathbf{m}).$$

Note that any $w(\mathbf{x_m}, \mathbf{m})$ that preserves $p(\mathbf{x})$ also preserves the joint distribution of feature and mask, i.e., $p_w(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y})$.

Proof. We only need to consider the MAR case, from which the MCAR case will follow. Let $w(\mathbf{x_m}, \mathbf{m}) = 0$ if $p(\mathbf{m} \mid \mathbf{x_m}) = 0$; otherwise, let

$$w(\mathbf{x_m}, \mathbf{m}) = \frac{1}{p(\mathbf{m} \mid \mathbf{x_m})} \prod_{i=1}^{n} \tilde{p}(m_i), \tag{1}$$

where $\tilde{p}$ can be any distribution of $\mathbf{m}$ with the entries m_i's mutually independent. The weighted training distribution is then

$$\begin{aligned} p_w(\mathbf{x}, \mathbf{m}) &= w(\mathbf{x_m}, \mathbf{m})p(\mathbf{x}, \mathbf{m}) \\ &= w(\mathbf{x_m}, \mathbf{m})p(\mathbf{x})p(\mathbf{m} \mid \mathbf{x_m}) = p(\mathbf{x}) \prod_{i=1}^{n} \tilde{p}(m_i), \end{aligned}$$

where the second equality follows from the MAR assumption. The desired results follow from marginalization. □

Despite the formula in Eqn. (1), we still need to actually learn $w(\mathbf{x_m}, \mathbf{m})$ from the data. To this end, we solve the following optimization problem that minimizes both intra-mask correlations and inter-correlations,

$$\min_{\mathbf{w} \in \mathbb{R}^+} \mathbb{WD} + \sum_{k,l} \mathrm{cor}(X_k, M_l, \mathbf{w}) + \sum_{k<l} \mathrm{cor}(M_k, M_l, \mathbf{w}) + \gamma \mathrm{CV}(\mathbf{w}).$$

Here $\mathbb{WD} = \mathbb{WD}(\mathcal{D}_w, \mathcal{D}')$ is the empirical Wasserstein distance between the weighted training set $\mathcal{D}_w$ and a new incomplete dataset $\mathcal{D}'$ generated as follows. We first learn $p(\mathbf{x})$ from $\mathcal{D}$ using MisGAN [16]. A sample of $\mathcal{D}'$ is then obtained by masking a complete feature sample drawn from the learned $p(\mathbf{x})$ using a mask sample drawn from $\mathcal{D}_w$, according to the MCAR mechanism. Ideally, minimizing $\mathbb{WD}$ enforces $p_w(\mathbf{x_m}, \mathbf{m}) = p_w(\mathbf{m})p(\mathbf{x_m})$, which, when $\mathbf{x}$ and $\mathbf{m}$ become independent, implies $p_w(\mathbf{x_m}) = p(\mathbf{x_m})$, and hence helps preserve $p(\mathbf{x})$.

In the case of MCAR, the weights in Eqn. (1) become independent of $\mathbf{x_m}$, i.e.

$$w(\mathbf{x_m}, \mathbf{m}) = w(\mathbf{m}) = \frac{1}{p(\mathbf{m})} \prod_{i=1}^{n} \tilde{p}(m_i).$$

Instead of the above more general approach, we can learn a function $w(\mathbf{m})$ directly by solving the following simpler problem,

$$\min_{\mathbf{w} \in \mathbb{R}^+} \sum_{1 \le k < l \le n} \mathrm{cor}(M_k, M_l, \mathbf{w}) + \gamma \mathrm{CV}(\mathbf{w}).$$

5. EXPERIMENT

5.1. Experimental Setup

We evaluate StableMiss+ on both synthetic and real-world datasets.

Gaussian-Mix. Following [3], we generate feature $\mathbf{X}$ from a Gaussian Mixture model, and the scalar label Y is linear to features.

House Sales. Following [3], we use dataset of house sales, which contains $n = 16$ features and a scalar house price as label.

MNIST [17]. Images of handwritten digits. Given an incomplete image, we aim to predict the corresponding complete image.

Traffic [4]. Average traffic speed within every hour from 1343 roads in the city of Chengdu, China, in 2018. Note that this dataset is naturally incomplete. Given incomplete history, we aim to predict the future traffic speed.

All the datasets except Traffic are complete. We generate incomplete datasets by imposing mask on the complete samples according to the following missing patterns. The missing level r is set from 10% to 90% at a step of 10%.

MCAR. We generate mask $\mathbf{M}$ that is independent of feature $\mathbf{X}$, but the entries of mask can be dependent. For each sample, its sample missing rate r_s has 80% to be r and 2.5% to be one of the other 8 levels respectively. In each sample, following [8], we generate a window of length $\lfloor n \cdot r_s \rfloor$ at a random position, where the $\lfloor n \cdot r_s \rfloor$ consecutive features in the window are missing.

Table 2: Performance on House Sales and MNIST datasets with MAR mask when trained under 50% missing level.

Method	Testing Missing Level									
	House Sales					MNIST				
	10%	**30%**	**50%**	**70%**	**90%**	**10%**	**30%**	**50%**	**70%**	**90%**
Partial VAE	44.22	42.3	41.35*	61.10	100.69	40.88	40.66	39.75*	58.28	95.33
MIWAE	47.99	47.26	46.62*	64.96	100.72	44.36	45.26	44.62*	61.85	95.36
P-BiGAN	53.76	48.18	49.25*	66.67	111.03	49.70	46.11	47.06*	63.44	104.89
NeuMiss	46.74	42.25	40.56*	60.49	101.22	43.21	40.62	39.02*	57.72	95.82
DWR	59.21	49.58	43.98*	66.85	119.65	54.74	47.40	42.19*	63.60	112.86
SRDO	50.31	43.25	39.77*	64.09	103.84	46.51	41.55	38.29*	61.05	98.25
StableNet	42.26	39.14	41.35*	55.19	95.15	39.06	37.75	39.75*	52.82	90.21
StableMiss	35.53	32.82	32.89*	49.54	83.06	31.60	31.13	31.93*	48.05	74.23
StableMiss+	**29.83**	**28.76**	**32.61***	**44.40**	**76.13**	**25.76**	**27.08**	**30.86***	**44.26**	**69.30**

Table 3: Performance on naturally incomplete Traffic dataset.

Partial VAE	MIWAE	P-BiGAN	NeuMiss	DWR	SRDO	StableNet	StableMiss	StableMiss+
16.08	14.82	15.79	15.10	15.93	15.72	13.97	11.84	**10.62**

MAR. Following [1], we generate feature-dependent mask, and the entries of mask can also be dependent. First, randomly selected 10% features are set to be observed in all the samples. The mask on the other features are generated depending on the selected features. The sample missing rate r_s is the missing proportion of the other 90% features, which is determined in the same way as MCAR.

We compare StableMiss+ with StableMiss as well as state-of-the-art methods for prediction with incomplete data, including NeuMiss, Partial VAE, MIWAE and P-BiGAN, and those for generalization under agnostic feature distribution shift, including DWR, SRDO and StableNet. We use Root Mean Square Error (RMSE) as metric.

5.2. Experiment Results

We evaluate StableMiss+ under various missing patterns and levels. The difference in missing level represents mask distribution shift.

Synthetic. We will show the performance of the optimal predictor. It is derived from the known feature distribution and label generation process, without which the performance cannot be reached. We use the gap to optimal to reflect generalizability. Table 1 shows the performance on Gaussian-Mix when trained under 50%. Due to space limit, only those tested under 10%, 30%, 50%, 70% and 90% are shown; the others are similar. StableMiss+ has the best performance, reducing gap to optimal by 8% in MCAR and 6% in MAR.

Real-World. Since the label generation process is unknown, the optimal predictor cannot be derived. We show the exact RMSE. Table 2 shows the performance on House Sales and MNIST when trained under 50%. Due to space limit, we only show the more complex MAR case. StableMiss+ has the best performance, reducing RMSE by 12% on House Sales and 12% on MNIST. Table 3 shows the performance on the naturally incomplete Traffic dataset, where the training and testing missing rate are 37% and 23%. Since the ground truth of missing values is unavailable, the metric is computed only on the observed entries. StableMiss+ also has the best performance, with 10% reduction in RMSE.

Ablation Study. We study the efficacy of preserving intra-feature correlations and removing inter-correlations in the case of MAR. We compare with two variants that do not preserve intra-feature correlations, one removing only intra-mask correlations and the other further removing inter-correlations. Table 4 shows the results on Gaussian-Mix when trained under 50%. Without preserving intra-feature correlations and/or removing inter-correlations, the performance drops by 32% and 27%. Moreover, without preserving infra-feature correlations, the performance is even worse than StableMiss that removes them, since the unpreserved infra-feature correlations may reversely mislead the prediction. We also study the influence of decorrelate in the way of MAR when in the case of MCAR. Table 5 shows the results on Gaussian-Mix when trained under 50%. The performance is close. For real-world data with unknown correlations, we can decorrelate in the way of MAR.

Table 4: Ablation study on Gaussian-Mix feature with MAR mask.

Method	**10%**	**30%**	**50%**	**70%**	**90%**
Intra-mask	683	460	334*	576	648
Intra-mask & Inter	601	428	328*	531	612
StableMiss	431	309	283*	406	468
StableMiss+	**409**	**285**	**278***	**381**	**455**

Table 5: Ablation study on Gaussian-Mix feature with MCAR mask.

Method	**10%**	**30%**	**50%**	**70%**	**90%**
MCAR	379	320	291*	309	370
MAR	382	326	301*	303	379

6. CONCLUSION

In this paper, we propose StableMiss+, a novel method for prediction with incomplete feature under agnostic mask distribution shift. We analyze the issue of removing the intra-feature correlations in StableMiss and address this issue by preserving the intra-feature correlations and simultaneously removing the intra-mask correlations and inter-correlations. Experiments on synthetic and real-world datasets show that StableMiss+ outperforms the state-of-the-art methods, with 10% reduction in RMSE.

7. REFERENCES

[1] M. Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux, "NeuMiss networks: Differentiable programming for supervised learning with missing values," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Virtual Event, Dec. 2020, p. 5980–5990.

[2] M. Morvan, J. Josse, E. Scornet, and G. Varoquaux, "What's a good imputation to predict with mmissing values?," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Virtual Event, Dec. 2021, pp. 11530–11540.

[3] Y. Zhu, J. Yuan, B. Jiang, T. Lin, H. Jin, X. Wang, and C. Zhou, "Prediction with incomplete data under agnostic mask distribution shift," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, Macao, China, Aug. 2023, pp. 4720–4728.

[4] B. Lu, X. Gan, W. Zhang, H. Yao, L. Fu, and X. Wang, "Spatio-temporal graph few-shot learning with cross-city knowledge transfer," in *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, Aug. 2022, pp. 1162–1172.

[5] Y. Wu, J. Ni, W. Cheng, B. Zong, D. Song, Z. Chen, Y. Liu, X. Zhang, H. Chen, and S. B. Davidson, "Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series," in *The 35th AAAI Conference on Artificial Intelligence*, Virtual Event, Feb. 2021, pp. 651–659.

[6] C. Ma, S. Tschiatschek, K. Palla, J. Hernandez-Lobato, S. Nowozin, and C. Zhang, "EDDI: Efficient dynamic discovery of high-value information with partial vae," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, June 2019, pp. 4234–4243.

[7] P. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete data sets," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, June 2019, pp. 4413–4423.

[8] S. Li and B. Marlin, "Learning from irregularly-sampled time series: A missing data perspective," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual Event, July 2020, pp. 5937–5946.

[9] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, "Causally regularized learning with agnostic data selection bias," in *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Republic of Korea, Oct. 2018, p. 411–419.

[10] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, UK, Aug. 2018, pp. 1617–1626.

[11] K. Kuang, R. Xiong, P. Cui, S. Athey, and B. Li, "Stable prediction with model misspecification and agnostic distribution shift," in *The 34th AAAI Conference on Artificial Intelligence*, New York, USA, Feb. 2020, pp. 4485–4492.

[12] Z. Shen, P. Cui, T. Zhang, and K. Kuang, "Stable learning via sample reweighting," in *The 34th AAAI Conference on Artificial Intelligence*, New York, USA, Feb. 2020, pp. 5692–5699.

[13] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, Virtual Event, June 2021, pp. 5372–5382.

[14] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., USA, 1986.

[15] E. Strobl, K. Zhang, and S. Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," *Journal of Causal Inference*, vol. 7, no. 1, 2019.

[16] S. Li, B. Jiang, and B. Marlin, "Learning from incomplete data with generative adversarial networks," in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, May 2019.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.