# Prediction with Incomplete Data under Agnostic Mask Distribution Shift*

**Yichen Zhu**[1], **Jian Yuan**[1], **Bo Jiang**[†1], **Tao Lin**[2], **Haiming Jin**[1], **Xinbing Wang**[1] and **Chenghu Zhou**[1]
[1]Shanghai Jiao Tong University
[2]Communication University of China
{zyc_ieee, yuanjian, bjiang}@sjtu.edu.cn, lintao@cuc.edu.cn, {jinhaiming, xwang8}@sjtu.edu.cn, zhouchsjtu@gmail.com

## Abstract

Data with missing values is ubiquitous in many applications. Recent years have witnessed increasing attention on prediction with only incomplete data consisting of observed features and a mask that indicates the missing pattern. Existing methods assume that the training and testing distributions are the same, which may be violated in real-world scenarios. In this paper, we consider prediction with incomplete data in the presence of distribution shift. We focus on the case where the underlying joint distribution of complete features and label is invariant, but the missing pattern, i.e., mask distribution may shift agnostically between training and testing. To achieve generalization, we leverage the observation that for each mask, there is an invariant optimal predictor. To avoid the exponential explosion when learning them separately, we approximate the optimal predictors jointly using a double parameterization technique. This has the undesirable side effect of allowing the learned predictors to rely on the intra-mask correlation and that between features and mask. We perform decorrelation to minimize this effect. Combining the techniques above, we propose a novel prediction method called StableMiss. Extensive experiments on both synthetic and real-world datasets show that StableMiss is robust and outperforms state-of-the-art methods under agnostic mask distribution shift.

## 1 Introduction

Data with missing values is ubiquitous in many applications due to sensor malfunction, incomplete sensing coverage, etc. Recent years have witnessed increasing attention on prediction with only incomplete feature, which consists of observed feature values and a mask that indicates which features are observed. Existing methods [Morvan *et al.*, 2020; Morvan *et al.*, 2021] assume that the training and testing distributions are the same. However, this assumption can be violated in real-world scenarios. In this paper, we study the problem of prediction with incomplete data in the presence of distribution shift. We focus on the scenario where the underlying joint distribution of complete features and label is invariant, but the missing pattern, i.e., mask distribution may be different between training and testing. Such mask distribution shift may result from different sensor deployment, data management, etc. For example, DiDi Traffic Speed [DiDiChuxing, 2018] is a naturally incomplete dataset throughout one year. The traffic network is almost unchanged, so we may reasonably assume the speed distribution is relatively stable. The average missing rate from Jan to Jun is 37% but drops to 23% from Jul to Dec, possibly because more and higher-quality sensors are deployed. Moreover, different from the transfer learning formulation [Pan and Yang, 2010], we assume the mask distribution shift is agnostic, since testing distribution is usually unavailable during training in practice.

Many methods for missing data can be used to predict with incomplete feature. Morvan *et al.* [2020] propose NeuMiss for linear regression with incomplete Gaussian feature. Morvan *et al.* [2021] then extend NeuMiss to nonlinear case by training it jointly with an MLP. Some missing data imputation methods [Yoon *et al.*, 2018; Li *et al.*, 2019; Ma *et al.*, 2019; Mattei and Frellsen, 2019; Li and Marlin, 2020] can also be used by treating the label as missing feature. However, none of these methods consider distribution shift. Their models learn the information of mask distribution and thus can hardly generalize under mask distribution shift.

Several methods have been proposed for prediction under agnostic feature distribution shift [Shen *et al.*, 2018; Kuang *et al.*, 2018; Kuang *et al.*, 2020; Shen *et al.*, 2020; Zhang *et al.*, 2021; Xu *et al.*, 2022]. It is assumed that the conditional label distribution given complete feature is invariant. They learn this invariant conditional distribution to achieve generalization. However, they are designed for complete data. In our setting, the conditional label distribution given trivially-imputed complete feature is not invariant, so these methods cannot be applied to incomplete data.

We observe that the conditional label distribution given observed feature values and mask is invariant between training and testing. As a result, there is an invariant optimal predictor for each mask respectively. Learning these optimal predictors can generalize under agnostic mask distribution shift. Since the number of optimal predictors increases exponentially with feature dimension [Morvan *et al.*, 2020], we approximate

*Extended version: https://arxiv.org/abs/2305.11197.
†Corresponding author.

the optimal predictors jointly using a double parameterization technique, i.e., we first parameterize the optimal predictors and then further parameterize the mapping from mask to the parameters of the optimal predictors. However, such parameterization has the undesirable side effect that the learned model may depend on the intra-mask correlation and the correlation between features and mask, since the training loss depends on these correlations. This may defy the generalizability of the learned model, as such correlations will change under mask distribution shift. Inspired by Xu *et al.* [2022], we decorrelate features and mask to make the learned model independent of their correlations and help it well approximate the optimal predictors. Combining all the techniques above, we have our StableMiss that can achieve generalization under agnostic mask distribution shift.

The contributions of this paper are summarized as follows.

- This paper proposes StableMiss, a novel method for prediction with incomplete data that is robust to agnostic mask distribution shift. To the best of our knowledge, this is the first method that considers agnostic mask distribution shift.
- Extensive experiments are conducted on both synthetic and real-world datasets. The results show that StableMiss is robust and outperforms the state-of-the-art methods under agnostic mask distribution shift.

## 2 Related Work

**Prediction with Incomplete Data.** Prediction with incomplete feature has attracted increasing attention recently. Morvan *et al.* [2020] derive the analytical expression of the optimal predictor for linear regression with incomplete Gaussian feature and proposes the NeuMiss network to approximate the optimal predictor. Morvan *et al.* [2021] then extend NeuMiss to nonlinear case by training it jointly with a Multi-Layer Perceptron (MLP). Besides these methods natively proposed for prediction, some missing data imputation methods can also be used by treating the label as missing feature. GAIN [Yoon *et al.*, 2018] is an adaptation of GAN [Goodfellow *et al.*, 2014], where the generator imputes missing entries, which the discriminator tries to distinguish from observed entries with partial information about the mask. MisGAN [Li *et al.*, 2019] learns the complete data distribution from incomplete data and uses it to supervise imputation. Partial VAE [Ma *et al.*, 2019], MIWAE [Mattei and Frellsen, 2019] and P-BiGAN [Li and Marlin, 2020] extend VAE [Kingma and Welling, 2014], IWAE [Burda *et al.*, 2016] and BiGAN [Donahue *et al.*, 2017] respectively to learn the prior and posterior distributions of incomplete feature given its latent representation and mask. However, all these methods assume that the training and testing distributions are the same. They can hardly generalize under distribution shift.

**Prediction under Agnostic Feature Distribution Shift.** Several methods have been proposed for prediction under agnostic feature distribution shift. They typically assume that the conditional label distribution given complete feature is invariant. Then they learn this invariant conditional distribution to achieve generalization by decorrelating the features. CRLR [Shen *et al.*, 2018] learns a weight for each sample respectively by minimizing the so-called confounder balancing loss, which is zero when the features are decorrelated. Then weighted Logistic regression is carried out with the learned weights. Based on CRLR, DGBR [Kuang *et al.*, 2018] also extracts nonlinear representations of features with deep autoencoder. DWR [Kuang *et al.*, 2020] learns a set of sample weights by minimizing the sum of covariance between pairs of features and then carries out weighted linear regression. SRDO [Shen *et al.*, 2020] first constructs a new dataset by sampling each feature independently from the training set to decorrelate among features and then learns sample weights by density ratio estimation [Sugiyama *et al.*, 2012]. Weighted linear regression is also used with the learned weights. StableNet [Zhang *et al.*, 2021] adopts Random Fourier Features to measure nonlinear correlations among features. It iteratively optimizes a weighted regression model and a set of sample weights by minimizing the prediction error and the nonlinear correlations, respectively. Xu *et al.* [2022] propose a general framework with DWR and SRDO as specific implementations and gives theoretical analysis on the framework. However, all these methods are designed for complete data and cannot be applied to incomplete data.

Compared to the existing methods, StableMiss can not only predict with incomplete feature but also generalize under agnostic mask distribution shift.

## 3 Problem Formulation

### 3.1 Preliminary

We use capital and lowercase letters, e.g., $X$ and $x$, to denote random variable and its realization, respectively. We use subscripts to 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. In the presence of missing data, we consider the case where $\mathbf{x}$ is partially observed and $\mathbf{y}$ is fully observed during training. A binary mask $\mathbf{m} \in \{0,1\}^n$ indicates 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 the 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., $p(\mathbf{y} \mid \mathbf{x}, \mathbf{m}) = p(\mathbf{y} \mid \mathbf{x})$. We do not make assumptions on $p(\mathbf{x})$.

Following Little and Rubin [1986], 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 resulted 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 Missing Completely At Random (MCAR) case and Missing At Random (MAR) case [Little and Rubin,

1986]. Under MCAR, the 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, $\mathbf{M}$ only depends on the observed feature values $\mathbf{X_M}$, i.e., $p(\mathbf{m} \mid \mathbf{x}) = p(\mathbf{m} \mid \mathbf{x_m}), \forall \mathbf{m}, \mathbf{x}$.

### 3.2 Problem Statement

The problem is to predict with incomplete feature under agnostic mask distribution shift. 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 $g(\mathbf{x_m}, \mathbf{m})$ for agnostic testing distribution $p^{te}(\mathbf{x}, \mathbf{m}, \mathbf{y})$, where the input to $g$ is only incomplete feature. We seek the prediction function $g$ to minimize the mean squared loss

$$\ell(g) = \mathbb{E}_{(\mathbf{X},\mathbf{M},\mathbf{Y})\sim p^{te}} \|\mathbf{Y} - g(\mathbf{X_M}, \mathbf{M})\|_2^2.$$

Ideally, the optimal $g$ is given by the conditional expectation:

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

How we learn it approximately from training data will be introduced in the next section.

Note that the testing error can be arbitrarily large without any prior knowledge about the testing distribution. We consider the case where only the mask distribution may change between training and testing. More specifically, we make the following assumption on the testing distribution.

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

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

The mask distribution shift still remains agnostic under the above assumption. Different from the transfer learning formulation [Pan and Yang, 2010], the testing distribution is unavailable during the training process.

## 4 Methodology

### 4.1 Prediction Framework

Our method relies critically on the following simple result, the proof of which is given in the extended version.

**Theorem 1.** *Under Assumption 1, the conditional label distribution given observed feature values and mask is invariant between training and testing in MCAR or MAR:*

$$p^{te}(\mathbf{y} \mid \mathbf{x_m}, \mathbf{m}) = p^{tr}(\mathbf{y} \mid \mathbf{x_m}, \mathbf{m}).$$

*As a consequence,*

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

Theorem 1 holds since MCAR and MAR guarantee that the missing feature $\mathbf{X}_{\overline{\mathbf{M}}}$ is independent of of mask $\mathbf{M}$ given the observed feature $\mathbf{X_M}$. It shows that the ideal optimal predictor is invariant between training and testing. If we can learn it under the training distribution, it will automatically generalize to the agnostic testing distribution. However, as noted in Morvan *et al.* [2021], it is essentially an aggregation of $2^n$ optimal predictors, one for each specific mask $\mathbf{m}$. Since the number of optimal predictors increases exponentially with feature dimension, it is infeasible to learn them separately.

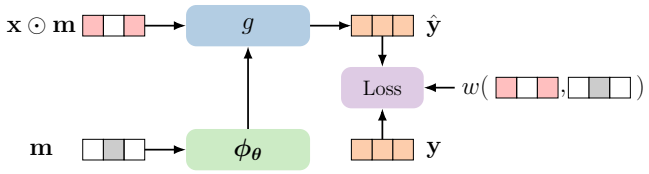


Figure 1: Prediction framework with sample reweighting.

Note that the optimal predictor can be thought of as a function of the observed feature $\mathbf{x_m}$ parameterized by $\boldsymbol{\phi}$, where $\boldsymbol{\phi}$ is a function of $\mathbf{m}$, i.e., $g(\mathbf{x_m}, \mathbf{m}) = g_{\boldsymbol{\phi}(\mathbf{m})}(\mathbf{x_m})$. Learning all the $2^n$ optimal predictors then corresponds to learning the $2^n$ different values of $\boldsymbol{\phi}(\mathbf{m})$. The function $\boldsymbol{\phi}(\mathbf{m})$ corresponding to the optimal $g$ is very complicated in general. To address the exponential explosion problem, we adopt the common technique of approximating it by a simpler function $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{m})$ parameterized by $\boldsymbol{\theta}$, which can be implemented by a neural network with parameter $\boldsymbol{\theta}$. Since we will also implement $g$ by a neural network, we use the zero-imputed feature instead of $\mathbf{x_m}$ to uniformize the input size. The final form of our predictor is thus

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

where $\odot$ is element-wise multiplication and $\boldsymbol{\theta}$ is the parameter we need to learn. The framework is shown in Figure 1.

We illustrate the above framework by a simple example. Suppose $Y = \sum_{i=1}^n \alpha_i X_i$, where $X_i$'s are mutually independent except that $X_1 = X_2$. The optimal predictor is

$$\mathbb{E}[Y \mid \mathbf{x_m}, \mathbf{m}] = \phi_0 + \sum_{i=1}^n \phi_i \cdot (\mathbf{x} \odot \mathbf{m})_i,$$

where

$$\begin{aligned} &\phi_0 = \bar{m}_1\bar{m}_2(\alpha_1\mathbb{E}X_1 + \alpha_2\mathbb{E}X_2) + \sum_{i=3}^n \bar{m}_i\alpha_i\mathbb{E}X_i, \\ &\phi_1 = (\alpha_1 + \alpha_2\bar{m}_2)m_1, \quad \phi_2 = (\alpha_2 + \alpha_1\bar{m}_1)m_2, \\ &\phi_i = \alpha_i m_i, \quad \text{for } i = 3, \ldots, n. \end{aligned} \tag{1}$$

Note that, if $x_i$ is missing, the value of $\phi_i$ actually does not matter, as it is multiplied by $(\mathbf{x} \odot \mathbf{m})_i = x_i m_i = 0$. Here the function $g$ is linear in $\mathbf{x} \odot \mathbf{m}$ with the parameter $\boldsymbol{\phi}(\mathbf{m}) = (\phi_0(\mathbf{m}), \ldots, \phi_n(\mathbf{m}))$ being quadratic in $\mathbf{m}$. We can easily parameterize $\boldsymbol{\phi}(\mathbf{m})$ by $\boldsymbol{\theta} \in \mathbb{R}^{(n+1)\times(n+1)\times(n+1)}$, where $\phi_k(\mathbf{m}) = \sum_{i,j} \theta_{ijk} m_i m_j$. The corresponding $g$ is then

$$g_{\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{m})}(\mathbf{x} \odot \mathbf{m}) = \sum_{i,j,k} \theta_{ijk} x_k m_i m_j m_k, \tag{2}$$

where $x_0 = m_0 = 1$. Here approximation is not necessary as the optimal $\boldsymbol{\phi}(\mathbf{m})$ is simple enough. In general, however, $\boldsymbol{\phi}(\mathbf{m})$ may be very complex and approximation is necessary. For example, when the $X_i$'s are not independent, the $\boldsymbol{\phi}(\mathbf{m})$ for the linear model can be an $n$-th order polynomial in $\mathbf{m}$ with $2^n$ parameters.

Note that the architecture of NeuMiss [Morvan *et al.*, 2020] fits into our framework, although it does not explicitly consider mask distribution shift. NeuMiss uses a specific $g$ and $\boldsymbol{\phi_\theta}$ that are carefully designed to approximate the analytical form of the optimal predictor for a linear label generation process with Gaussian feature. Our framework is more flexible and applies to more general settings, as shown in Section 5.

### 4.2 Decorrelation

The use of $\boldsymbol{\phi_\theta}(\mathbf{m})$ solves the exponential explosion problem, but it introduces another. As noted in prior work [Shen *et al.*, 2018; Kuang *et al.*, 2018; Kuang *et al.*, 2020; Shen *et al.*, 2020; Zhang *et al.*, 2021; Xu *et al.*, 2022], the learned value of $\boldsymbol{\theta}$ can be affected by the correlation between the variables that we condition on in the optimal predictor, i.e., $\mathbf{X}$ and $\mathbf{M}$ in our case. This defies our original goal, as a learned value of $\boldsymbol{\theta}$ that performs well on the training distribution may not do so on the testing distribution, where the correlation changes. We explain this problem in details below.

We assume $\mathbb{E}[\mathbf{Y} \mid \mathbf{x_m}, \mathbf{m}]$ is realizable, i.e., there exists a parameter $\boldsymbol{\theta}^*$ corresponding to the parameter of optimal predictor $\boldsymbol{\phi}$. In the above example, this assumption is satisfied by the form of $g$ in Equation (2), and $\boldsymbol{\theta}^*$ corresponds to the function $\boldsymbol{\phi}$ as given by Equation (1). In practice, we use neural networks for $g$ to approximately satisfy this assumption, which shows good empirical results in Section 5.

Consider the population loss. $\boldsymbol{\theta}^*$ minimizes training loss

$$\ell_{tr}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},\mathbf{M},\mathbf{Y})\sim p^{tr}} \|\mathbf{Y} - g_{\boldsymbol{\phi_\theta}(\mathbf{M})}(\mathbf{X_M}, \mathbf{M})\|_2^2.$$

Since parameter $\boldsymbol{\theta}^*$ corresponds to the parameter of optimal predictor, it also minimizes the above loss under testing distribution $p^{te}$, i.e., $\boldsymbol{\theta}^*$ generalizes under mask distribution shift.

Suppose $\hat{\boldsymbol{\theta}}$ is the parameter learned from training distribution, i.e. $\ell_{tr}(\hat{\boldsymbol{\theta}}) = \ell_{tr}(\boldsymbol{\theta}^*)$. $\hat{\boldsymbol{\theta}}$ can be different from $\boldsymbol{\theta}^*$, but they perform equally well on the training distribution. In fact,

$$g_{\boldsymbol{\phi}_{\hat{\boldsymbol{\theta}}}(\mathbf{M})}(\mathbf{X_M}, \mathbf{M}) = g_{\boldsymbol{\phi}_{\boldsymbol{\theta}^*}(\mathbf{M})}(\mathbf{X_M}, \mathbf{M}), \quad p^{tr}\text{-a.s..} \tag{3}$$

However, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ may not perform equally well on testing distribution $p^{tr}$, as formally stated in Theorem 2, the proof of which is given in the extended version.

**Theorem 2.** *In the realizable case, there exists a problem instance, in which there exists a $\hat{\boldsymbol{\theta}}$ other than $\boldsymbol{\theta}^*$ such that Equation* (3) *holds $p^{tr}$-a.s. but not $p^{te}$-a.s..*

Such $\hat{\boldsymbol{\theta}}$ is undesirable, since it cannot generalize as well as $\boldsymbol{\theta}^*$. To address this problem, decorrelation among $\mathbf{X}$ and $\mathbf{M}$ may help. We explain the reason with an example below.

Consider the example in Section 4.1. If the entries of $\mathbf{X}$ and the entries of $\mathbf{M}$ are mutually independent, the only solution for $\hat{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}^*$ corresponding to Equation (1). The proof is given in the extended version. Thus we can perform decorrelation to help obtain the desired parameter $\boldsymbol{\theta}^*$.

Under MCAR, we decorrelate the entries of $\mathbf{X}$ and those of $\mathbf{M}$ respectively; under MAR, we further decorrelate between $\mathbf{X}$ and $\mathbf{M}$. In this paper, we follow Zhang *et al.* [2021] to decorrelate by sample reweighting. Specifically, we learn a weighting function $w(\mathbf{x_m}, \mathbf{m})$ by minimizing the correlation under distribution $\tilde{p}_w$, where

$$\tilde{p}_w(\mathbf{x}, \mathbf{m}, y) = w(\mathbf{x_m}, \mathbf{m}) p^{tr}(\mathbf{x}, \mathbf{m}, y).$$

The learned weighting function will be normalized to make $\tilde{p}_w$ well-defined. Under $\tilde{p}_w$, the conditional label distribution $p(\mathbf{y} \mid \mathbf{x_m}, \mathbf{m})$ does not change. Inspired by Zhang *et al.* [2021], we measure the correlation empirically by partial cross-covariance matrix with Random Fourier Feature [Strobl *et al.*, 2019]. We adapt it to the case of incomplete data as follows.

The partial cross-covariance matrix of $X_k$ and $X_l$, denoted by $\Sigma'_{X_k,X_l;w}$, is computed with only samples in which $X_k$ and $X_l$ are both observed:

$$\Sigma'_{X_k,X_l;w} = \frac{1}{N^{kl}-1} \sum_{i=1}^{N^{kl}} \Bigg[ \Bigg( w_i \mathbf{u}(X_k^{(i)}) - \frac{1}{N^k} \sum_{j=1}^{N^k} w_j \mathbf{u}(X_k^{(j)}) \Bigg)^\top \Bigg( w_i \mathbf{v}(X_l^{(i)}) - \frac{1}{N^l} \sum_{j=1}^{N^l} w_j \mathbf{v}(X_l^{(j)}) \Bigg) \Bigg],$$

where $N^{kl}$ is the number of samples in which $X_k$ and $X_l$ are both observed, $N^k$ and $N^l$ are the number of samples with observed $X_k$ and $X_l$, respectively, $w_i = w(\mathbf{x}^{(i)}_{\mathbf{m}^{(i)}}, \mathbf{m}^{(i)})$, and $\mathbf{u}$, $\mathbf{v}$ are function vectors with elements drawn from the space of Random Fourier Feature [Strobl *et al.*, 2019], the definition of which is given in the extended version. Similarly, $\Sigma'_{X_k,M_l;w}$ is computed with only samples in which $X_k$ is observed, and $\Sigma'_{M_k,M_l;w}$ is computed with all the $N$ samples. See the extended version for their detailed forms.

When decorrelating the entries of $\mathbf{X}$ and those of $\mathbf{M}$, we optimize sample weight $\mathbf{w} \in \mathbb{R}_+^N$ by minimizing the correlation of all pairs of $X_i$ and $X_j$ and all pairs of $M_i$ and $M_j$:

$$\min_{\mathbf{w}\in\mathbb{R}^+} \sum_{1\le k<l\le n} \|\Sigma'_{X_k,X_l;w}\|_F^2 + \sum_{1\le k<l\le n} \|\Sigma'_{M_k,M_l;w}\|_F^2 + \gamma \frac{\mathrm{Std}(\mathbf{w})}{1/N\sum_{i=1}^N w_i},$$

where $\mathrm{Std}(\mathbf{w})$ is standard deviation of the $w_i$'s. The last regularization term is used to prevent assigning very large weights to a small proportion of samples. Hyper-parameter $\gamma$ is coefficient for the regularization term. Under MAR, we further add $\sum_{1\le k,l\le n} \|\Sigma'_{X_k,M_l;w}\|_F^2$ into the objective to decorrelate between $\mathbf{X}$ and $\mathbf{M}$.

With the learned weighting function, we can conduct regression under the weighted distribution $\tilde{p}_w$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\tilde{p}_w}[(Y - g_{\boldsymbol{\phi_\theta}(\mathbf{M})}(\mathbf{X_M}, \mathbf{M}))^2].$$

In the ideal case, the entries of $\mathbf{X}$ and those of $\mathbf{M}$ are mutually independent under $\tilde{p}_w$. Then the learned $\boldsymbol{\theta}$ will not be affected by the correlation and can generalize under mask distribution shift. In practice, we can equivalently conduct weighted regression under the training distribution:

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

How decorrelation is combined with the prediction framework is shown in Figure 1.

## 5 Experiment

### 5.1 Datasets

We evaluate StableMiss on synthetic and real-world datasets.

| Method | | Testing Missing Level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **10%** | **20%** | **30%** | **40%** | **50%** | **60%** | **70%** | **80%** | **90%** |
| **Gap to Optimal** | Partial VAE | 1541.52 | 1244.90 | 1172.68 | 961.18 | 793.32* | 898.04 | 1075.69 | 1087.19 | 1112.42 |
| | MIWAE | 1331.95 | 1029.18 | 1047.14 | 910.77 | 743.63* | 841.89 | 820.39 | 844.80 | 844.17 |
| | P-BiGAN | 1382.71 | 1123.64 | 1041.63 | 923.53 | 713.85* | 856.39 | 845.39 | 867.93 | 979.45 |
| | NeuMiss | <u>714.19</u> | <u>520.26</u> | <u>523.10</u> | <u>432.99</u> | <u>293.14</u>* | <u>574.80</u> | <u>601.74</u> | <u>646.53</u> | <u>679.99</u> |
| | DWR | 1485.42 | 1170.15 | 1142.32 | 1033.32 | 893.64* | 962.61 | 969.56 | 966.85 | 963.52 |
| | SRDO | 1385.00 | 1116.82 | 1080.93 | 988.86 | 843.85* | 919.26 | 899.21 | 903.82 | 866.09 |
| | StableNet | 1256.32 | 967.22 | 989.67 | 878.97 | 743.25* | 818.16 | 807.50 | 750.89 | 788.56 |
| | StableMiss | **431.16** | **342.70** | **309.09** | **278.97** | **282.75*** | **319.58** | **405.60** | **412.03** | **467.98** |
| **Optimal** | | 904.87 | 1096.36 | 1108.73 | 1224.50 | 1324.00 | 1375.16 | 1518.59 | 1626.74 | 1705.54 |

Table 1: Performance on Gaussian-Mix feature with MAR mask when trained under 50% missing level. The values for Optimal are the 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 <u>underline</u> represent the best and second best along each column, respectively. Superscript * indicates when training and testing missing levels are the same. These marks will also be used in the other tables.

**Gaussian**. Following Morvan *et al.* [2021], we generate feature $\mathbf{X}$ from multivariate Gaussian distribution. The mean values are drawn from standard Gaussian distribution, and the covariance matrix is generated by $\Sigma = \mathbf{B}\mathbf{B}^\top + \mathbf{D}$, where the entries of $\mathbf{B} \in \mathbb{R}^{n\times 0.7n}$ are drawn from standard Gaussian distribution and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is diagonal with values uniformly drawn from $[10^{-2}, 10^{-1}]$. The feature dimension $n = 50$.

**Gaussian-Ind**. The feature generation is the same as Gaussian, except that we make the entries of $\mathbf{X}$ mutually independent by using a diagonal covariance matrix.

**Gaussian-Mix**. We generate feature $\mathbf{X}$ from the more general Gaussian mixture model with 3 components, each generated in the same way as Gaussian. The proportion of the $i$-th component $\pi_i$ is uniformly drawn from $[0, 1)$ and normalized by $\pi_i / \sum_{i=1}^{3} \pi_i$.

For the above 3 synthetic features, the label generation process is linear: $Y = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i + \epsilon$, where $\epsilon$ is a Gaussian noise such that the signal-to-noise ratio is 10. Different from Morvan *et al.* [2020; 2021] where all $\alpha_i$'s except $\alpha_0$ are equal, we draw more general $\alpha_i$'s from Gaussian distribution.

**House Sales**. Following Shen *et al.* [2020], we use dataset of house sales in King County, USA, which contains $n = 16$ features and a scalar house price as label.

**MNIST** [Lecun *et al.*, 1998]. The MNIST dataset of handwritten digit images. Given incomplete image, we aim to predict the complete image.

**Traffic** [DiDiChuxing, 2018]. Average traffic speed within every hour from 1343 roads in the city of Chengdu, China, in 2018. We build a graph for the dataset, where nodes represent the roads and edges indicate the adjacency of roads. Note that this dataset is naturally incomplete. Given incomplete history, we aim to predict the future traffic speed.

All the datasets except the Traffic dataset are complete. We generate incomplete datasets by imposing mask on the complete samples according to the missing patterns in Section 5.2.

## 5.2 Missing Patterns

We design the following 3 missing patterns. There are 9 missing levels, denoted by $r$, from 10% to 90% at a step of 10%. Missing patterns examples are given in the extended version.

**MCAR-Ind**. We generate mask $\mathbf{M}$ that is independent of feature $\mathbf{X}$, and the entries of mask are mutually independent. 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. Each entry is independently missing with probability $r_s$.

**MCAR**. We generate mask $\mathbf{M}$ that is independent of feature $\mathbf{X}$, but the entries of mask can be dependent. The sample missing rate $r_s$ is determined in the same way as MCAR-Ind. In each sample, following Li and Marlin [2020], 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.

**MAR**. Following Morvan *et al.* [2020; 2021], 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 according to a model whose parameters depend on the selected features. The sample missing rate $r_s$ means the missing proportion of the other 90% features, which is determined in the same way as MCAR-Ind.

## 5.3 Baselines

We compare StableMiss with two categories of baselines. (1) state-of-the-art methods on prediction with incomplete data: NeuMiss [Morvan *et al.*, 2020; Morvan *et al.*, 2021], Partial VAE [Ma *et al.*, 2019], MIWAE [Mattei and Frellsen, 2019] and P-BiGAN [Li and Marlin, 2020]; (2) state-of-the-art methods on generalization under agnostic feature distribution shift: DWR [Kuang *et al.*, 2020], SRDO [Shen *et al.*, 2020] and StableNet [Zhang *et al.*, 2021]. For Partial VAE, MIWAE and P-BiGAN, we treat label as missing feature. For DWR, SRDO and StableNet, we use mean-imputed feature as input. See the extended version for implementation details.

We use the commonly adopted Root Mean Square Error (RMSE) as evaluation metric throughout.

## 5.4 Performance on Synthetic Dataset

We compare StableMiss with baselines under different missing patterns and missing levels. The models are trained under a specific missing level and tested under all the missing

| Method | | Training Missing Level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **10%** | **20%** | **30%** | **40%** | **50%** | **60%** | **70%** | **80%** | **90%** |
| **Gap to Optimal** | Partial VAE | 1441.03 | 946.28 | 951.77 | 880.44 | 793.32* | 1048.26 | 1581.51 | 1608.35 | 1897.15 |
| | MIWAE | 1092.03 | 806.06 | 816.18 | 753.13 | 743.63* | 883.19 | 1333.60 | 1235.24 | 1674.31 |
| | P-BiGAN | 1136.98 | 840.27 | 849.73 | 771.86 | 713.85* | 947.65 | 1215.25 | 1255.63 | 1709.70 |
| | NeuMiss | 687.57 | 650.05 | 529.37 | 424.46 | 293.14* | 692.26 | 744.01 | 850.52 | 1023.01 |
| | DWR | 1111.11 | 957.66 | 962.37 | 904.77 | 893.64* | 1054.67 | 1312.80 | 1434.32 | 1902.80 |
| | SRDO | 991.91 | 939.84 | 890.83 | 884.84 | 843.85* | 1029.93 | 1362.18 | 1359.29 | 1897.69 |
| | StableNet | 856.08 | 847.25 | 866.97 | 823.89 | 743.25* | 889.24 | 1336.49 | 1410.46 | 1697.40 |
| | StableMiss | **337.14** | **326.49** | **296.93** | **297.95** | **282.75*** | **421.15** | **480.80** | **584.96** | **716.63** |
| **Optimal** | | | | | | 1324.00 | | | | |

Table 2: Performance on Gaussian-Mix feature with MAR mask when tested under 50% missing level.

| Method | | Testing Missing Level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **10%** | **20%** | **30%** | **40%** | **50%** | **60%** | **70%** | **80%** | **90%** |
| **Gap to StableMiss-ID** | Partial VAE | 23.37 | 18.94 | 17.94 | 14.85 | 8.46* | 19.70 | 21.87 | 46.05 | 52.53 |
| | MIWAE | 27.14 | 23.92 | 22.90 | 21.16 | 13.73* | 21.52 | 25.73 | 49.95 | 52.56 |
| | P-BiGAN | 32.91 | 25.62 | 23.82 | 23.10 | 16.36* | 26.29 | 27.44 | 56.17 | 62.87 |
| | NeuMiss | 25.89 | 20.19 | 17.89 | 14.48 | 7.67* | 17.67 | 21.26 | 46.55 | 53.06 |
| | DWR | 38.36 | 26.80 | 25.22 | 23.70 | 11.09* | 26.23 | 27.62 | 61.13 | 71.49 |
| | SRDO | 29.46 | 22.49 | 18.89 | 15.16 | 6.88* | 15.65 | 24.86 | 51.77 | 55.68 |
| | StableNet | 21.41 | 15.18 | 14.78 | 14.53 | 8.46* | 18.55 | 15.96 | 38.09 | 46.99 |
| | StableMiss | **14.68** | **10.16** | **8.46** | **7.03** | **0.00*** | **9.51** | **10.31** | **21.90** | **34.90** |
| **StableMiss-ID** | | 20.85 | 22.80 | 24.36 | 28.77 | 32.89 | 36.67 | 39.23 | 42.90 | 48.16 |

Table 3: Performance on House Sales dataset with MAR mask when trained under 50% missing level (unit: $10000).

levels. The difference in missing level represents the mask distribution shift. Due to space limit, we only present the results of missing rate shift on Gaussian-Mix feature with MAR mask; see the extended version for the missing rate shift on other settings and missing pattern shift. We also present the performance of optimal predictor $\mathbb{E}[Y \mid \mathbf{x_m}, \mathbf{m}]$. It can be derived from known feature distribution and label generation process, without which the performance of optimal predictor is not reachable.

We will show the results from two views: results with fixed training missing level that reflects the generalizability of methods, and results with fixed testing missing level that reflects the robustness to the mask distribution of training data.

**Fixing Training Missing Level**. We show the results when trained under 50% missing level and tested under all the missing levels in Table 1; the other settings are similar. Since the amount of observed data is different between missing levels, which influences the prediction error, we use the gap to optimal to reflect the generalization performance.

As the missing level increases, the error of the optimal predictor also increases, as there are less observed data. When the mask distribution shifts, StableMiss has the best generalization performance, reducing the gap to optimal of the second best, NeuMiss in this case, by 31%-44%. When there is no mask distribution shift, StableMiss still has the best performance in this case, slightly outperforming NeuMiss by 4% in terms of the gap to optimal. As shown in the extended version, for Gaussian features with MCAR or MAR mask, NeuMiss can slightly outperform StableMiss, by at most 2%, in the absence of mask distribution shift. This is not too surprising, as these are the cases that NeuMiss is tailored for.

**Fixing Testing Missing Level**. We show the results when tested under 50% missing level and trained under all the missing levels in Table 2; the other settings are similar. When the mask distribution shifts, StableMiss achieves the best performance, reducing the gap to optimal of the second best NeuMiss by 30%-50%. Reduction is especially large when trained under 10% missing level, since the optimal predictors are more learnable under lower missing level. The results show that StableMiss is robust to the quality, i.e., mask distribution, of training data and can generalize from various training mask distributions.

## 5.5 Performance on Real-World Dataset

**House Sales**. We show the results on House Sales dataset when trained under 50% missing level in Table 3. Without prior knowledge about label generation process, the optimal predictor cannot be derived. StableMiss has the best performance when trained in distribution, i.e., trained under testing distribution. Instead of Optimal, we show our in-distribution performance, which is denoted by 'StableMiss-ID'.

Error of StableMiss-ID also increases with missing level.

| Method | | Testing Missing Level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **10%** | **20%** | **30%** | **40%** | **50%** | **60%** | **70%** | **80%** | **90%** |
| **Gap to StableMiss-ID** | Partial VAE | 21.61 | 17.51 | 16.58 | 13.74 | 7.82* | 18.22 | 20.22 | 42.58 | 48.57 |
| | MIWAE | 25.09 | 22.12 | 21.18 | 19.57 | 12.69* | 19.90 | 23.79 | 46.18 | 48.60 |
| | P-BiGAN | 30.43 | 23.69 | 22.03 | 21.36 | 15.13* | 24.31 | 25.38 | 51.94 | 58.13 |
| | NeuMiss | 23.94 | 18.67 | 16.54 | 13.39 | 7.09* | 16.34 | 19.66 | 43.04 | 49.06 |
| | DWR | 35.47 | 24.78 | 23.32 | 21.92 | 10.26* | 24.25 | 25.54 | 56.52 | 66.10 |
| | SRDO | 27.24 | 20.79 | 17.47 | 14.01 | 6.36* | 14.47 | 22.99 | 47.87 | 51.49 |
| | StableNet | 19.79 | 14.03 | 13.67 | 13.43 | 7.82* | 17.15 | 14.76 | 35.22 | 43.45 |
| | StableMiss | **12.33** | **8.75** | **7.05** | **6.65** | **0.00*** | **8.25** | **9.99** | **18.73** | **27.47** |
| **StableMiss-ID** | | 19.27 | 22.14 | 24.08 | 27.96 | 31.93 | 34.95 | 38.06 | 41.65 | 46.76 |

Table 4: Performance on MNIST dataset with MAR mask when trained under 50% missing level.

| Partial VAE | MIWAE | P-BiGAN | NeuMiss | DWR | SRDO | StableNet | StableMiss | StableMiss-ID |
|---|---|---|---|---|---|---|---|---|
| 16.08 | 14.82 | 15.79 | 15.10 | 15.93 | 15.72 | 13.97 | **11.84** | 8.39 |

Table 5: Performance on Traffic dataset. All the values are exact RMSE (unit: km/h).

When the mask distribution shifts, StableMiss achieves the best performance, reducing the gap to StableMiss-ID of the second best, usually StableNet, by 26%-50%, with an average of 38%, which demonstrates our generalizability on real-world data.

**MNIST**. We show the results on MNIST dataset when trained under 50% missing level in Table 4. When the mask distribution shifts, StableMiss achieves the best performance, reducing the gap to StableMiss-ID of the second best, usually StableNet, by 28%-50%, with an average of 41%.

**Traffic**. Table 5 shows the results on Traffic dataset, which is naturally incomplete. The training and testing missing rates are 37% and 23%. Since ground truth of missing entries is unavailable, RMSE is only computed on observed entries. The value for StableMiss-ID is the training error, which only serves as a reference for our comparison. Note that StableMiss achieves the best performance, reducing the gap to StableMiss-ID of the second best, StableNet, by 38%, which shows that StableMiss can be applied to real-world datasets with complex missingness.

### 5.6 Ablation Study

We study the efficacy of decorrelation by comparing with ablated variants that only decorrelates the entries of $\mathbf{X}$ and $\mathbf{M}$ respectively (only intra), only decorrelates between $\mathbf{X}$ and $\mathbf{M}$ (only inter) and does not decorrelate (w/o). The result on Gaussian-Mix feature with MAR mask when trained under 10% missing level is given in Figure 2. Our method with full decorrelation achieves the best performance especially when mask distribution shift becomes stronger. The gap of the second best to optimal is reduced by 55% on average and by 83% under 90% missing level.

We also study the influence of decorrelation when it is not needed. The result on Gaussian-Ind feature with MCAR-Ind mask when trained under 10% missing level is given in Figure 3. The performance with unneeded decorrelation is close to that without decorrelation. For real-world data with unknown correlation, we can always conduct decorrelation.

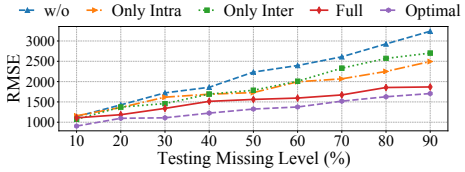


Figure 2: Comparison with ablated variants: only decorrelating the entries of $\mathbf{X}$ and $\mathbf{M}$ respectively (only intra), only decorrelating between $\mathbf{X}$ and $\mathbf{M}$ (only inter), and no decorrelaion (w/o). The same notations are used in Figure 3.

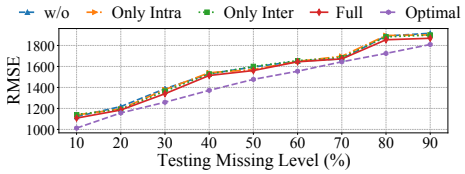


Figure 3: Influence of unneeded decorrelation.

## 6 Conclusion

In this paper, we propose a novel method StableMiss for the problem of *prediction with incomplete data under agnostic mask distribution shift*. We leverage the observation that for each mask, there is an invariant optimal predictor. We approximate the optimal predictors jointly using a double parameterization technique. We also perform decorrelation to minimize the side effect caused by the intra-mask correlation and the correlation between features and mask. Extensive experiments are conducted on both synthetic and real-world datasets. The results show that StableMiss is robust and outperforms state-of-the-art methods under agnostic mask distribution shift.

## Acknowledgments

This work was supported by NSF China (No. 62072302, U20A20181, U21A20519, 42050105, 62262018) and the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (No.SKLMCC2021KF011).

## References

[Burda *et al.*, 2016] Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *The 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

[DiDiChuxing, 2018] DiDiChuxing. City traffic index. https://gaia.didichuxing.com/, 2018. Accessed: 2018-12-31.

[Donahue *et al.*, 2017] Jeff Donahue, Philipp Krahenbuhl, and Trevor Darrell. Adversarial feature learning. In *The 5th International Conference on Learning Representations*, Toulon, France, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, Montreal, Canada, December 2014.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*, Banff, Canada, 2014.

[Kuang *et al.*, 2018] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1617–1626, London, UK, 2018.

[Kuang *et al.*, 2020] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *The 34th AAAI Conference on Artificial Intelligence*, pages 4485–4492, New York, USA, 2020.

[Lecun *et al.*, 1998] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li and Marlin, 2020] Steven Cheng-Xian Li and Benjamin Marlin. Learning from irregularly-sampled time series: A missing data perspective. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5937–5946, Virtual Event, 2020.

[Li *et al.*, 2019] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Learning from incomplete data with generative adversarial networks. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, May 2019.

[Little and Rubin, 1986] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., USA, 1986.

[Ma *et al.*, 2019] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial vae. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4234–4243, Long Beach, USA, 2019.

[Mattei and Frellsen, 2019] Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4413–4423, Long Beach, USA, 2019.

[Morvan *et al.*, 2020] Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gael Varoquaux. NeuMiss networks: Differentiable programming for supervised learning with missing values. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, page 5980–5990, Vancouver, Canada, 2020.

[Morvan *et al.*, 2021] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with mmissing values? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 11530–11540, Virtual Event, 2021.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Shen *et al.*, 2018] Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM International Conference on Multimedia*, page 411–419, Seoul, Republic of Korea, 2018.

[Shen *et al.*, 2020] Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kuang. Stable learning via sample reweighting. In *The 34th AAAI Conference on Artificial Intelligence*, pages 5692–5699, New York, USA, 2020.

[Strobl *et al.*, 2019] Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

[Sugiyama *et al.*, 2012] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, UK, 2012.

[Xu *et al.*, 2022] Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, and Peng Cui. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 24803–24829, Baltimore, USA, 2022.

[Yoon *et al.*, 2018] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5675–5684, Stockholmsmassan, Sweden, July 2018.

[Zhang *et al.*, 2021] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, Virtual Event, 2021.