Spectral Distribution Alignment for Enhanced Generalization in Regression

Kaiyu Guo[1], Zijian Wang[1], Brian C. Lovell[2], and Mahsa Baktashmotlagh[1] (✉)

[1] University of Queensland, Brisbane, Australia, 4072 {uqkguo1, zijian.wang, m.baktashmotlagh}@uq.edu.au

[2] University of Queensland, Brisbane, Australia, 4072 lovell@itee.uq.edu.au

Abstract. While several techniques have been proposed to enhance the generalization of deep learning models for classification problems, limited research has been conducted on improving generalization for regression tasks. This is primarily due to the continuous nature of regression labels, which makes it challenging to directly apply classification-based techniques to regression tasks. In this paper, we introduce a novel generalization method for regression tasks based on the metric learning assumption that the distance between features and labels should be proportional. Unlike previous approaches that solely consider the prediction of this proportion as constant and disregard its variation among samples, we argue that this proportion can be defined as a mapping function. Additionally, we propose minimizing the error of this function and stabilizing its fluctuating behavior by smoothing out its variations. To further enhance Out-of-Distribution (OOD) generalization, we leverage the characteristics of the spectral norm (*i.e.*, the sub-multiplicativity of the spectral norm of the feature matrix can be expressed as Frobenius norm of the output), and align the maximum singular value of the feature matrices across different domains. We conduct experiments on 5 datasets for OOD generalization in regression, and our method consistently outperforms state-of-the-art approaches in the majority of cases. Our code is released at https://github.com/workerbcd/SCR

Keywords: Out-of-Distribution Generalization · Representation learning · Regression .

1 Introduction

Continuous label prediction, known as regression, finds widespread application across various domains, including computer vision [49,8], medical testing [18,1], and financial analysis [22]. Unlike classification, which aims to determine optimal decision boundaries, regression involves fitting outputs to a continuous function [30]. On the other hand, out-of-distribution generalization has received considerable attention in classification [40]; however, the exploration of regression generalization remains relatively limited. Specifically, existing representation learning-based methods [4,2,16,17,14] are predominantly tailored for classification tasks. Although these methods can be adapted for regression general-

ization, their efficacy is constrained as they fail to fully account for the inherent proportional interdependence between features and labels.

In light of the above discussion, we argue that, when addressing challenges such as uncertainty estimation [25] and generalization [46] in regression, it is crucial to consider the relationships between the labels. To this end, we introduce a metric learning loss specifically designed for regression. Different from the previous method RankSim[20] to regularize the distribution with the order of label distance, our proposed loss not only takes the distance order into account but also constrains the feature distribution by discriminating the ratio between feature distance and label distance. This idea can efficiently solve the discontinuity of feature distribution caused by the discrete ranking order in RankSim. In addition, contrary to the assumption in Regression Metric Loss (RML) [7] that the ratio between feature distance and label distance is constant, we show that this ratio varies and only equals a constant under certain ideal conditions. We argue that RML, by overlooking the variability in this ratio, may obscure the pattern of feature distributions in certain cases. As demonstrated in Figure 1, our metric loss can significantly discriminate the pattern, while maintaining the continuity of feature distribution.

To improve the OOD generalization, we design a method to align the discriminative feature pattern in different distributions. Motivated by augmentation-based techniques [44,39,48] for OOD generalization in classification, we further leverage the method mixing pairs of training data [46] to generate new distributions. For each distribution, we create a metric penalty to identify discriminative patterns within the feature distribution. The real and synthesized distributions are aligned by minimizing the difference between the spectral norms of their feature representations. According to the spectral norm property, the minimization process ensures that the Frobenius norm of outputs remains consistent, thereby reducing the upper bound of distribution discrepancy in regression tasks. The main contributions of this paper are three-folded:

- We introduce a tailored metric loss for regression, bringing features closer if their label distances are small and pushing them apart if distances are large. This facilitates pattern recognition in regression and improves generalization performance. We depart from previous approaches by modeling the feature-label distance proportion as a variable mapping function and mitigating instability caused by its fluctuations.
- According to the theory of domain adaptation in regression, we theoretically present the relation between the spectral norm of feature matrix and the upper bound of the distribution discrepancy in regression. Based on our theory, we expand the training distribution by generating samples with new distribution and then align the real and synthesized distributions by minimizing the difference between the spectral norm of their feature representations, to enhance OOD generalization in regression.
- We conduct experiments on five regression datasets and show that our method outperforms the state-of-the-art in most cases. The t-SNE visualization of

the feature embedding illustrates the effectiveness and stability of our proposed metric loss.

2 Related Work

2.1 Metric Learning

Metric learning has been shown to be effective when related to methods that rely on distances and similarities [28]. Traditionally, methods like PCA and KNN are widely used in the area of machine learning. With the development of deep learning, networks [38] related to pair distances are designed to correlate among samples while using shared weights in deep learning [26]. Then, prototype-based metric losses [41,13] were proposed based on contrastive motivation. In regression tasks, the metric learning loss has not been well-defined because it is hard to build the connection between the metric distance and continuous labels. Recently, RML [7] has been proposed based on the assumption that there is a constant proportion between the feature distance and the label distance. However, the method based on this assumption only considers the scale of the feature matrix, ignoring fluctuations in the proportion map. To solve this issue, we assume that the proportion is a mapping function in the training process and propose a metric loss to smooth fluctuations.

2.2 Out-of-Distribution Generalization

Out-of-distribution (OOD) generalization aims at generalizing the model from the training distribution to an unseen distribution. Mostly, the methods can be divided into 3 parts [40]: data augmentation, representation learning, and training strategy. Data augmentation methods [48,50] utilize linear interpolation to fill the distribution gap, and some methods [44,39] also generate a new distribution to enrich the convex hull supported by the source distributions. Representation learning [4,3] aims at generating distribution-invariant feature representations from source distributions. Recently, methods like SWAD [6] proposed some novel training and model selection strategies, significantly improving performance in OOD generalization. However, most methods above are designed for classification. There are limited methods [46,4] designed for regression tasks.

2.3 Generalization in Regression

Recent research targeting generalization in regression tasks is based on data augmentation in which mixup pairs are selected based on the probability related to label distances [46,45]. Even though limited research has been proposed on this topic, some methods designed for regression tasks can be transferred to generalization purposes. For instance, due to the function of metric learning, the metric loss in regression [20,47,7] can be regarded as an in-distribution generalization method. Also, distribution alignment methods in regression [34,11] can be updated as OOD generalization methods. However, these distribution alignment

methods are not related to the proportion between features and labels, which are supposed to be very important in regression tasks.

3 Methodology

3.1 Problem Definition

Let $\{(x_i, y_i)\}_{i=1}^{N}$ be the dataset with N samples, with $x_i \in \mathcal{X}$ being the input sample $i \in \mathbb{R}^+$ and $y_i \in \mathcal{Y}$ its corresponding label, and $\mathcal{X}$ and $\mathcal{Y}$ denoting the input space and the continuous label space, respectively. In the training phase, the network learns a projection function $g : \mathcal{X} \to \mathcal{F}$ and a regression function $\varphi : \mathcal{F} \to \mathcal{Y}$. The projection function g transforms the input data into the feature space, and the regression function φ maps the compact feature representation to the label space. The objective of the regressor is to bring the output prediction $\hat{y}_i$ close to the ground truth label y_i. Ideally, the optimal predictor φ is a fully connected layer that satisfies $y_i = \hat{y}_i = W_\varphi^* f_i + b_\varphi^*$, where $f_i = g(x_i)$ is the extracted feature, W_φ^* is the optimal weight, and b_φ^* is the optimal bias.

Distribution discrepancy in regression. A theory of learning from different distributions in regression is defined in [12]. Given the hypothesis h being a map from input space $\mathcal{X}$ to the label space $\mathcal{Y}$, the discrepancy distance $disc$ between two distributions P and Q is defined as:

$$disc(P, Q) = \max_{h, h' \in H} |\mathcal{L}_P(h', h) - \mathcal{L}_Q(h', h)|$$

Here, the hypothesis H is a subspace of the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ and $\mathcal{L}_D(h', h) = E_{x \sim D}[L(h(x), h'(x))]$, with L being a MSE loss. In this paper, we only consider the situation of finite dimension, thus, Euclidean space can be considered as a Hilbert space with a linear kernel.

3.2 Proportional Metric Loss

By leveraging the discrete labels to define sample pairs in classification models, metric learning aims to learn feature representations with low intra-class variance and high inter-class separation, which can improve the generalization ability of the learned model [28,9]. However, this motivation is based on the fact that the labels are discrete [34]. In regression tasks, given an input-label pair of (x_i, y_i), $\forall \epsilon > 0$, with input $x_{i+\epsilon}$ and its continuous label $y_{i+\epsilon}$, it's proven that φ should be a continuous bijection [7], with homeomorphic label and feature distributions. Intuitively, there is an optimal relationship between the distances of labels and distances of features - as the distance between two labels increases, the distance between their corresponding features should also increase, meaning that when two examples have labels that are farther apart, their representations in feature space should also be farther apart, and vice versa for labels that are closer together.

Remark 1 $d(y_i, y_j) < d(y_i, y_k) \iff d(f_i, f_j) < d(f_i, f_k), \forall i, j, k \in \mathbb{R}^+$

For any bounded open subset in $\mathcal{F}$, φ should be convergent and bounded, which means φ should be uniformly continuous on any bounded open subset [36]. Thus, we conclude Remark 2.

Remark 2 $d(y_i, y_j) < d(y_t, y_k) \Longleftrightarrow d(f_i, f_j) < d(f_t, f_k), \forall i, j, k, t \in \mathbb{R}^+$

Building upon Remark 1, since $\mathcal{F}$ is a compact space and label $\mathcal{Y}$ is continuous, then for $\forall \epsilon > 0$, we can find labels y', y'' with $d(y', y'') = \epsilon$. Then, $\exists \delta = d(f', f'') > 0$, such that $\forall d(f_a, f_b) < \delta$, we have $d(y_a, y_b) < \epsilon$. So, Remark 2 keeps φ uniformly continuous.

In light of the discussion above, we argue that the distance between labels can not be ignored in the regression tasks. In particular, we propose learning a feature-label proportional distance instead of the traditional distance, *e.g.*, Euclidean distance between features:

$$d_r(f_i, f_j) = \frac{d(f_i, f_j)}{d(y_i, y_j)}, \tag{1}$$

Here, $d(\cdot, \cdot)$ represents Euclidean distance and $d_r(\cdot, \cdot)$ denotes the proportional distance induced from $d(\cdot, \cdot)$. In addition, $d_r(\cdot, \cdot)$ should be a bounded distance, which can be illustrated by the following Proposition.

Proposition 1 *Given any two data points (x_i, y_i) and (x_j, y_j), we have $\|f_i - f_j\|_p \leq \|{W^*}_\varphi^{-1}\|_p |y_i - y_j|$. Here, W_φ^* is the optimal weight of the fully connected layer. f_i, f_j are the features extracted from x_i, x_j through model g, and $\|\cdot\|_p$ is the norm under L_p space.*

Proof 1 *Given the optimal weight W_φ^*, bias b_φ^* and data (x_i, y_i), (x_j, y_j), we have*

$$y_i = W_\varphi^* f_i + b_\varphi^*, y_j = W_\varphi^* f_j + b_\varphi^*$$

where f_i, f_j are extracted features from x_i, x_j, respectively. Then,

$$\|f_i - f_j\|_p = \|W_\varphi^{*\dagger}(y_i - y_j)\|_p \leq \|W_\varphi^{*\dagger}\|_p |y_i - y_j|$$

where $\dagger$ represents Moore–Penrose inverse

Proposition 1 gives the upper bound of $d_r(\cdot, \cdot)$ which is $\|W_\varphi^{*\dagger}\|_2$. In addition, when the equal sign in Proposition 1 holds, it can explain the assumption of regression metric loss [7] that the distance between the features should be proportional to the distance between their corresponding labels. Specifically, [7] uses a learnable parameter to restrain the proportion between feature distance and label distance. However, according to Proposition 1, this proportion should be related to the optimal weight W_φ^*, and the equation may not hold when the labels are continuous. Moreover, representing this proportion with a constant ignores its fluctuations and variances among different samples. To alleviate this issue, we formulate this proportion as a mapping function and minimize its standard deviation to constrain the distance between the features to be uniform along the samples.

According to Proposition 1, the result of $d_r(\cdot,\cdot)$ should be a bounded proportion map and can be a constant function in some ideal situation. Hence, we minimize the standard deviation of $d_r(\cdot,\cdot)$ to acquire a flatter proportion map in a mini-batch. The proportional metric loss function should be:

$$L_{pml} = \sqrt{\frac{1}{N_b^2 - 1}\sum_{i}^{N_b}\sum_{j}^{N_b}(d_r(f_i, f_j) - \bar{d}_r)} \tag{2}$$

Here, $\bar{d}_r$ is a constant function equal to the mean of the relative distances in the batch and N_b is the batch size. L_{pml} constrains the predictor φ as a Lipschitz continuous function satisfying Remarks 1 and 2. L_{pml} is based on the assumption that the target label is univariate. For the multivariate regression task with a D dimensional target label $y \in \mathbb{R}^D$, the loss can be updated as:

$$\hat{L}_{pml} = \sum_{i=1}^{D} L_{pml}^{i} \tag{3}$$

with $L_{pml}^{i} = L_{pml}$ if $D = 1$.

Superiority of L_{pml} over SOTA RankSim [20] is the SOTA method to regularize the feature distance in regression by aligning the feature distance order with the label distance order. We argue that the consistency between the orders of label distance and feature distances is insufficient to regularize the feature distribution with continuous labels, especially for unseen labels.[3] With the consideration of the proportion between label distance and feature distance, L_{pml} can mitigate this problem with continuous feature distribution.

3.3 Spectral Alignment of Domains

Existing works [44,43] in domain generalization have demonstrated that the diversity and amount of training examples are positively correlated with the generalizability of a machine learning model. To expand the training set, we employ the data augmentation technique of C-Mixup [46] to generate additional samples from unseen distributions. However, without imposing a constraint of domain invariance, the learned feature space might include domain-specific information and thus become noisy [32]. This could hinder obtaining the optimal generalization power of the model.

To impose domain invariance constraint, the existing work of [11] suggests not to minimize the difference between the Frobenius norm of feature representations of different domains. However, this may cause unstable performance.

[3] In the 1-D space, given seen labels $l_s = \{1, 2, 10, 11\}$, the feature can be distributed as $f_s = \{1, 2, 100, 101\}$ according to [20]. But, according to the pigeonhole principle, there must be at least two unseen labels in $[2, 10]$ with label distance equal to 0.5 (default to $\{4, 4.5\}$), whose feature distance is larger than 1 (default to $\{6, 7.1\}$). However, the features $f = \{1, 2, 6, 7.1, 100, 101\}$ with labels $l = \{1, 2, 4, 4.5, 10, 11\}$ are ill-distributed according to [20].

We argue that this instability can come from the fact that the Frobenius norm may encode the average of variances (i.e., singular values) along all orthogonal feature projections, and that, the transferability of the feature representations mainly lies in aligning the highest variability directions corresponding to the largest singular values [10]. Therefore, in our formulation, the Frobenius norm is substituted by the spectral norm, which only encodes the highest variability direction. We further show that the difference between spectral norms of features can be related to domain discrepancy.

Notations The expected loss in regression is $\mathcal{L}_D(h', h) = E_{x\sim D}[L(h(x), h'(x))]$ with L being the MSE loss [12]. We have the $\mathcal{L}_D(h, 0) = \frac{1}{N}\|\hat{Y}_D^h\|_F^2$, with N being the number of samples, and $\hat{Y}_D^h$ being the output with hypothesis h under distribution D. 0 represents the hypothesis mapping to zero element in $\mathcal{Y}$.

Proposition 2 *Given two distributions P and Q, we have*

$$disc(P, Q) \leq \frac{1}{N} \max_{h\in H} |\|\hat{Y}_P^h\|_F^2 - \|\hat{Y}_Q^h\|_F^2|,$$

where disc represents the difference between distributions and N denotes the number of the samples.

Proof 2 *Generally speaking, we have*

$$\mathcal{L}(h', h) = \mathcal{L}(h - h', 0)$$

Since h, h' are in the subspace H of Hilbert Space $\mathbb{H}$, we have $h'' = h - h' \in H$. Then, we have

$$\forall h'' \in \mathbb{H}, disc(P, Q) \leq \max_{h''\in H} |\mathcal{L}_P(h'', 0) - \mathcal{L}_Q(h'', 0)|$$

So, the proof is concluded.

Proposition 2 shows the relation between the difference of feature representations and their distribution discrepancy. To determine the relation between the norm of the feature matrix and the output scale[4], we consider the spectral norm of the feature space, $\|F\|_2 = \sup_{w\neq 0} \frac{\|Fw\|_2}{\|w\|_2}$. If W_i is a row vector of the weight W in the fully connected layer, then $\|\hat{Y}_i^h\|_2 \leq \|\hat{Y}_i^h - b_i\|_2 + |b_i| \leq \|F\|_2\|W_i\|_2 + |b_i|$, $\hat{Y}_i^h$, and $\hat{Y}_i^h$ is the i-th vector of the output matrix $\hat{Y}^h$ and b_i is the i-th value of the bias vector b in the fully connected layer. If we define $\lambda_i(F) = \|F\|_2\|W_i\|_2 + |b_i|$, we will have $\|\hat{Y}^h\|_F \leq \|\lambda(F)\|_2$.

From the discussion above, the spectral norm is related to the upper bound of the output scale. So aligning the spectral norms can prevent the output scales from differing greatly, which can also align two distributions as per Proposition

[4] The Frobenius norm of the output $\|\hat{Y}_P^h\|_F$ represents the scale of the output in distribution P. Unlike classification, in regression, the target label for each sample can be a vector. That means, if we have N samples, each with M dimensional target vectors, then $\hat{Y}_P^h$ is an $N \times M$ matrix.

2. Therefore, we propose the spectral alignment loss based on singular value decomposition (SVD) as follows:

$$L_{sa} = |max(s^{real}) - max(s^{syn})|, \tag{4}$$

where s_{real} and s_{syn} are the set of the singular values of the feature matrices from the real and synthesized distributions. The largest singular values of matrices are selected for calculating the loss. Note that $\|F\|_2 = max(s_F)$, where s_F is the set of the singular values of matrix F.

3.4 Overall Objective Function

We combine our objectives for proportional metric loss and spectral alignment, and optimize them in an end-to-end training fashion. Formally, we have:

$$L = L_{mse} + \alpha \hat{L}_{pml} + \beta L_{sa}, \tag{5}$$

where α and β represent hyper-parameters to balance the contribution of their corresponding loss functions. We further optimize the supervised loss of L_{mse}, formulated as:

$$L_{mse} = \frac{1}{N}(\sum_{i=1}^{N}(\varphi(g(x_i^{real})) - y_i^{real})^2 + \sum_{i=1}^{N}(\varphi(g(x_i^{syn})) - y_i^{syn})^2) \tag{6}$$

with $\varphi(g(x_i^{real}))$ and $\varphi(g(x_i^{syn}))$ being the prediction of input x_i^{real} and the augmented sample x_i^{syn} with C-Mixup [46], respectively. Here, y_i^{real} and y_i^{syn} denote the ground truth label corresponding to x_i^{real} and x_i^{syn}, respectively.

4 Experimental Results

4.1 Implementation Details

Recent research [29,27] reveals a phenomenon that fine-tuning the whole network on a new task can improve the in-distribution (ID) performance of the new task, at the price of its out-of-distribution (OOD) accuracies. This is because fine-tuning the whole network changes the feature space spanned by the training data of a new task, which distorts the pretrained features. While linear probing can be an alternative solution to fine-tuning, due to its inability to adapt the features to the downstream task, it may degenerate the performance on in-distribution tasks. To mitigate this ID-OOD trade-off, motivated by the discussion in [29,27], we freeze the top of the C-Mixup [46] pretrained network (excluding the last block and the linear layers) during the training process. Specifically, we only fine-tune the bottom layer to preserve the low-level features from the pretrained model and unfreeze the last block to avoid degeneracy in the ID tasks.

Table 1: Comparison on out-of-distribution datasets. The **bold** number is the best result and the underline number is the second best result. The results of methods with † are reported by [46]. ERM is the baseline method using only MSE loss. The results of methods with * are reproduced based on the provided source code

	RCF-MNIST	DTI	
	RMSE↓	R↑	
	Avg.	Avg.	Worst
ERM†	0.162	0.464	0.429
ERM*	0.160	0.475	0.438
IRM† [4] †	0.153	0.478	0.432
IB-IRM† [2]	0.167	0.479	0.435
CORAL† [31]	0.163	0.483	0.432
GroupDRO† [37]	0.232	0.442	0.407
Mixup† [48]	0.176	0.465	0.437
C-Mixup* [46]	0.153	0.483	0.449
C-Mixup† [46]	0.146	<u>0.498</u>	0.458
RML* [7]	0.167	0.480	0.446
RankSim* [20]	0.239	0.479	<u>0.464</u>
Full model w/o L_{sa}	<u>0.145</u>	0.491	**0.479**
Full model w/o L_{pml}	0.147	0.481	0.447
Full model	**0.143**	**0.500**	0.448

4.2 Generalization in Univariate Regression

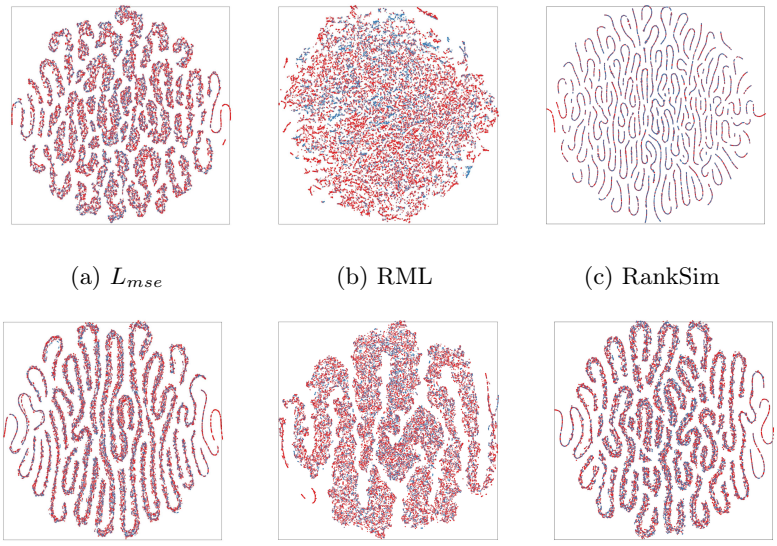
Datasets. The generalization ability of models in regression tasks with univariate output is evaluated over two datasets, namely Drug-target Interactions (DTI) [24] and RCF-MNIST [46]. **DTI** is a real world dataset designed to predict the binding activity score between each small molecule and the corresponding target protein by collecting 232,458 data on the drug and target protein information. The whole dataset is divided into different domains according to the years of data collection. **RCF-MNIST** is a dataset with 60,000 images built on FashionMNIST [42] with spurious correlations between colours and rotation angles (label).

Experimental Settings. We evaluate our method on two datasets, namely RCF-MNIST and DTI. We leverage Resnet18 [23] as the feature extractor for RCF-MNIST, and employ DeepDTA [35] on DTI.

Following the original paper of DTI [24], we evaluate the methods on R value. Same as C-Mixup [46], we report both average and worst-domain performance for the experiments on DTI. For RCF-MNIST, the evaluation metric is Root Mean Square Error (RMSE). Our full model is trained with three losses, L_{mse}, L_{sa} and L_{pml}. The fine-tuning strategy mentioned in Section 4.1 is also applied in the experiments on univariate regression in our models. All the experiments are run over 3 seeds.

Performance comparison. The performance of OOD robustness on the two datasets is shown in Table 1. We compare our methods with not only the OOD

generalization methods, *i.e* C-Mixup [46], CORAL [31], but also some metric loss in regression, *i.e* RankSim [20] and RML [7]. Note that RankSim is a method designed for age prediction where the continuity of the target label is not required. ERM is the baseline training strategy, where the objective is to minimize the Mean Squared Error (MSE) loss Similar to our proposed losses, the fine-tuning method is applied for the models with RankSim and RML. As the table shows, our method can achieve superior performance in most cases. For the datasets with small sizes, the pretrained model plays an important role in improving generalization, since the scarcity of data and the lack of variety is the key problem in these datasets. In addition, we find that L_{pml} can also generalize the spurious correlation, as shown by the results of RCF-MNIST. We assume that the spurious correlation increases the variance in the proportion, which can be generalized by L_{pml}.



(a) L_{mse} (b) RML (c) RankSim

(d) Full model w/o L_{sa} (e) Full model w/o L_{pml} (f) Full model

Fig. 1: T-SNE visualization of the embedding space on DTI dataset. The visualizations from le to right are (a) The baseline model that is fine-tuned to minimize MSE loss, (b) The model that is fine-tuned to minimize both MSE and RML objectives, (c) the model that is fine-tuned to minimize both MSE and RankSim, (d) The model that is fine-tuned to optimize full model without L_{sa}, (e) The model that is fine-tuned to optimize full model without L_{pml}. (f) The model that is fine-tuned to optimize full model. Red points represent the features extracted from the train set and the blue points represent the features extracted from the test set.

t-SNE visualization. According to our discussion, L_{pml} is trying to get a flatter d_r, which means the feature distribution should follow a discriminative pattern with less variance. To test the effect of the losses in regression on embedding space, we visualize the feature distribution without metric loss, with RML, and with L_{pml} on Figure 1. This visualization can strongly support our assumption and discussion above. As Figure 1 shows, the feature distribution is more compact and the distribution pattern is clearer with L_{pml}. In addition, as we discussed, RML focuses on learning a scale of the matrix feature and ignores the variance in the proportion. So, in some situations, the pattern will be blurred with RML, which is the same as the one shown in Figure 1. Note that L_{pml} maintains the property of being Lipschitz continuous for the predictor, which enhances the continuity of the feature distribution with less steep slopes. Figures 1c and 1d illustrate this difference: unlike L_{pml}, RankSim [20], which focuses solely on the distance between orders, does not preserve Lipschitz continuity. This characteristic might contribute to L_{pml}'s superior performance over RankSim in most scenarios, as shown in Tables 1. It will also contribute to the frequent breakpoints in Figure 1c, which supports this hypothesis.
L_{pml} *v.s. Ranksim*: The t-SNE visualization highlights two primary distinctions between the methods: the pattern of L_{pml} appears rougher than that of RankSim; and L_{pml} exhibits significantly fewer breakpoints compared to RankSim. This is likely due to the penalty mechanism of RankSim, which aligns feature distances more loosely in accordance with label distances, allowing for a broader spread with fewer disruptions. This accounts for the rougher appearance of the L_{pml} pattern. However, such stretching of patterns might result in extremely varied feature distances, potentially causing poorly distributed patterns, particularly for unseen labels as noted in Footnote 3. Consequently, the t-SNE visualization of RankSim reveals more breakpoints than that of L_{pml}, explaining the suboptimal performance of RankSim in DTI and RCF-MNIST.

4.3 Generalization in Multivariate Regression

Datasets. The out-of-distribution (OOD) generalization ability of models in multivariate regression is evaluated over three datasets, including dSprites [33], MPI3D [19] and BiwiKinect [15] which are widely used for domain adaptation tasks in computer vision [11,34]

dSprites is a synthetic dataset of three domains, namely Color(**c**), Scream(**s**) and Noisy(**n**), which are generated by adding colors or noise in the real images. Each domain comprises 737,280 images. Following the setup in [11], the orientation factors in the dataset are excluded.

MPI3D is a benchmark dataset of 1,036,800 images with three distributions to predict intrinsic factors. The dataset contains real data (domain Real (**rl**)) and synthetic data (domain Toy (**t**) and Realistic (**rc**)). In our experiments, we only consider the prediction of the rotation around a vertical and horizontal axis.

BiwiKinect is a real-world dataset of head poses recorded by a Microsoft Kinect sensor. The dataset can be divided into 2 domains: Female (**F**) with 5,874

Table 2: Comparison on MPI3D and dSprites dataset with the setting of domain generalization under the MSE index. The **bold** number is the best and the underline number is the second best result. The unseen domains are on the top.

	MPI3D-MSE			MPI3D-MAE		
	rc	rl	t	rc	rl	t
ERM	0.08132	0.09819	0.007004	0.3163	0.3511	0.0922
C-Mixup	0.09226	0.10495	0.014453	0.3367	0.3666	0.1296
RML	0.08596	0.09412	0.020132	0.3315	0.3448	0.1661
Nuclear-norm	0.09490	0.09536	0.011940	0.3270	0.3313	0.1181
F-norm	0.09565	0.10548	0.008318	0.3226	0.3411	0.0985
Full model w/oL_{sa}	**0.07829**	0.08262	0.006996	0.3149	0.3478	0.0919
Full model w/o L_{pml}	0.07942	0.08355	**0.006016**	**0.3016**	0.3225	**0.0856**
Full model	0.07956	**0.07885**	0.006017	0.3058	**0.3137**	0.0858
	dSprites-MSE			dSprites-MAE		
	c	s	n	c	s	n
ERM	0.04904	0.4903	0.4108	0.3071	0.9793	0.8977
C-Mixup	0.08769	0.5087	**0.3672**	0.4144	0.9846	**0.8596**
RML	0.08037	0.5860	0.4348	0.4147	1.054	0.9115
Nuclear-norm	0.2076	0.7718	0.4970	0.3270	0.3313	0.1181
F-norm	0.06709	0.4856	0.5868	0.3035	0.9234	1.067
Full model w/o L_{sa}	0.03480	0.4589	0.4051	0.2626	0.9413	0.8953
Full model w/o L_{pml}	0.03721	0.4861	0.3954	0.2711	0.9650	0.8859
Full model	**0.03479**	**0.4569**	0.4030	**0.2620**	**0.9396**	0.8942

images and Male (**M**) with 9,804 images. The Euler angles of the head, namely yaw, pitch and roll angles are used to evaluate our method.

Experimental settings. We analyze our method under the setting of domain generalization on the three datasets, which is a benchmark dataset for Domain Adaptation in Regression. We adapt a domain generalization setting [21] by evaluating our method over three generalization tasks on the three datasets: 1) dSprites:**c, s** $\rightarrow$ **n**; **c, n** $\rightarrow$ **s**; **s, n** $\rightarrow$ **c**; 2) MPI3D: **rl, rc** $\rightarrow$ **t**; **t, rc** $\rightarrow$ **rl**; **rl, t** $\rightarrow$ **rc**; 3) Biwikinect: **F** $\rightarrow$ **M**; **M** $\rightarrow$ **F**. We use the test sets of source distributions as the validation sets for the model selection. All the experiments are run over three random seeds, and we follow [6] for random seed and hyper-parameter seed selection.

The evaluation metrics on this task are Mean Square Error (MSE) and Mean Absolute Error (MAE). The variances over seeds of the methods are reported in the supplementary. We do not use our fine-tuning method on the three datasets and there is no frozen parameter during the training process on these three datasets. Since RankSim does not provide the algorithm for multivariate regression, we do not include it in the experiments for multivariate regression.

OOD generalization with multiple source domains. The MSE and MAE results on dSprites and MPI3D are shown in Table 2. The comparison between L_{pml} and RML [7] shows the advantage of modeling the proportion as a fluctuating map rather than a fixed constant. In addition, the performance also shows

Table 3: Comparison on BiwiKinect dataset with the setting of domain generalization under MSE and MAE index. The **bold** number is the best and the underline number is the second best result. The unseen domains are on the top.

	BiwiKinect-MSE		BiwiKinect-MAE	
	F	M	F	M
ERM	0.3953	0.4949	0.7907	0.8879
C-Mixup	0.3542	0.4908	0.7555	0.8795
RML	0.4139	0.4923	0.8125	0.8833
Nuclear-norm	0.4792	0.5967	0.8771	0.9902
F-norm	0.3472	0.4735	0.7394	0.8489
Full model w/o L_{sa}	0.3486	0.4744	0.7401	0.8424
Full model w/o L_{pml}	**0.3376**	**0.4683**	**0.7257**	0.8585
Full model	0.3391	0.4695	0.7276	**0.8419**

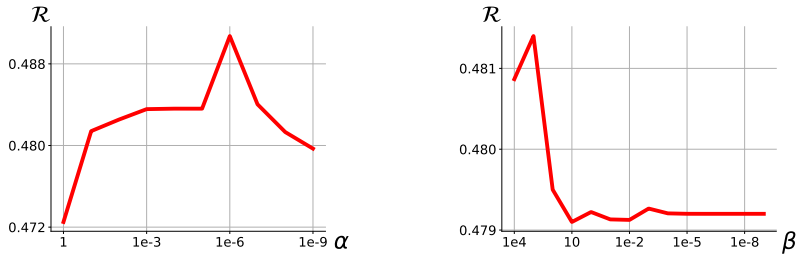
that the alignment with L_{sa} can significantly improve the generalization ability in some cases. We also provide the results of alignment with Nuclear-norm $\|\cdot\|_*$ and Frobenius norm $\|\cdot\|_F$. With norm equivalence [5], $\|\cdot\|_2 \leq \|\cdot\|_F \leq \|\cdot\|_*$, the spectral norm can give a tighter upper bound. This can explain the reason that L_{sa} can get the best performance among them. In addition, compared with L_{pml}, L_{sa} makes more significant improvement in generalization tasks, which illustrates the importance of distribution alignment in OOD generalization.

Table 2 shows that the C-Mixup performs very well, when the Noisy domain is the unseen domain on dSprites. We assume that the mix-up distribution is very similar to the noisy domain. When the other distributions are involved, the well-built distribution is damaged, which may complicate the training process.

OOD generalization with single source domain. We also evaluate our method on single OOD generalization on BiwiKinect dataset. Table 3 shows the MSE and MAE results on BiwiKinect. From the result of F-norm and L_{sa}, it seems that the distribution alignment methods can contribute more to the improvement of the performance on single domain generalization because of the lack of diversity in the source distribution.

4.4 Hyper-parameter Sensitivity Analysis

We analyze the hyper-parameters on α and β in Equation 5. When the values of L_{pml} and L_{sa} are much larger than L_{mse}, the total loss L_{mse} is hard to converge and the performance will drop dramatically. So, we analyze the trend of the performance of L_{pml} and L_{sa} with α and β in the range between $[1e^{-9}, 1]$ and $[1e^{-9}, 1e^{4}]$ respectively. Figure 2 shows the sensitivity of the hyper-parameters on out-of-distribution dataset DTI. We find that the L_{pml} is much more sensitive since the value of L_{pml} is usually much larger than L_{mse} and L_{sa}.



(a) α on DTI (b) β on DTI

Fig. 2: (a) and (b) shows the hyper-parameter analysis on DTI datasets. The larger $\mathcal{R}$ value means the better result.

5 Conclusion

This paper discusses two main objectives that are required to improve generalization in regression. For In-Distribution generalization, we propose proportional metric loss, based on the assumption that the distance between features and their corresponding labels should be correlated. We assume that the proportion between feature distance and label distance is a mapping function. Through this loss, we show that the variance in the embedding space is decreased, resulting in more discriminative patterns. To improve the transferability of the model on out-of-distribution data, we propose to augment the original data and then align the synthesized and real distributions through minimizing the difference between spectral norm of features.

Acknowledgments. This work is partially supported by Australian Research Council Project FT230100426.

References

1. Agatston, A., Janowitz, W., Hildner, F.J., Zusmer, N.R., Viamonte, M., Detrano, R.: Quantification of coronary artery calcium using ultrafast computed tomography. Journal of the American College of Cardiology **15**(4), 827–832 (1990) 1
2. Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.C., Bengio, Y., Mitliagkas, I., Rish, I.: Invariance principle meets information bottleneck for out-of-distribution generalization. Proceeding of the Conference on Neural Information Processing Systems **34**, 3438–3450 (2021) 1, 9
3. Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T.H., Mitliagkas, I.: Generalizing to unseen domains via distribution matching. arXiv:1911.00804 (2019) 3
4. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019) 1, 3, 9
5. Cai, T.T., Ren, Z., Zhou, H.H.: Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation (2016) 13

6. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Proceeding of the Conference on Neural Information Processing Systems **34**, 22405–22418 (2021) 3, 12
7. Chao, H., Zhang, J., Yan, P.: Regression metric loss: Learning a semantic representation space for medical images. arXiv preprint arXiv:2207.05231 (2022) 2, 3, 4, 5, 9, 10, 12
8. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Proceeding of the Conference on Neural Information Processing Systems (2016) 1
9. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2017) 4
10. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1081–1090. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/chen19i.html 7
11. Chen, X., Wang, S., Wang, J., Long, M.: Representation subspace distance for domain adaptation regression. In: International conference on machine learning. pp. 1749–1759 (2021) 3, 6, 11
12. Cortes, C., Mohri, M.: Domain adaptation in regression. In: International Conference on Algorithmic Learning Theory. pp. 308–323. Springer (2011) 4, 7
13. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019) 3
14. Deshmukh, A.A., Lei, Y., Sharma, S., Dogan, U., Cutler, J.W., Scott, C.: A generalization error bound for multi-class domain generalization. arXiv preprint arXiv:1905.10392 (2019) 1
15. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. International Journal of Computer Vision **101**(3), 437–458 (2013). https://doi.org/10.1007/s11263-012-0549-0 11
16. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015) 1
17. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016) 1
18. Gilsanz, V., Ratib, O.: Hand Bone Age: A Digital Atlas of Skeletal Maturity. Springer Berlin Heidelberg (2011) 1
19. Gondal, M.W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., Bauer, S.: On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) 11
20. Gong, Y., Mori, G., Tung, F.: Ranksim: Ranking similarity regularization for deep imbalanced regression. arXiv preprint arXiv:2205.15236 (2022) 2, 3, 6, 9, 10, 11
21. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2021) 12
22. Happersberger, D.: Advancing Systematic and Factor Investing Strategies Using Alternative Data and Machine Learning. Lancaster University (2021) 1

23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2016) 9
24. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C.W., Xiao, C., Sun, J., Zitnik, M.: Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv:2102.09548 (2021) 9
25. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**, 457–506 (2021) 2
26. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. Symmetry **11**(9), 1066 (2019) 3
27. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=Zb6c8A-Fghk 8
28. Kulis, B., et al.: Metric learning: A survey. Foundations and Trends® in Machine Learning **5**(4), 287–364 (2013) 3, 4
29. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054 (2022) 8
30. Lee, C., Landgrebe, D.A.: Feature extraction based on decision boundaries. Transactions on Pattern Analysis and Machine Intelligence **15**(4), 388–400 (1993) 1
31. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5400–5409 (2018) 9, 10
32. Liu, Y., Wang, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3510–3519 (2023) 6
33. Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017) 11
34. Nejjar, I., Wang, Q., Fink, O.: Dare-gram : Unsupervised domain adaptation regression by aligning inversed gram matrices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2023) 3, 4, 11
35. Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics **34**(17), i821–i829 (2018) 9
36. Rudin, W.: Principles of Mathematical Analysis. International series in pure and applied mathematics, McGraw-Hill (1976) 5
37. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019) 9
38. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–823 (2015) 3
39. Sicilia, A., Zhao, X., Hwang, S.J.: Domain adversarial neural networks for domain generalization: When it works and how to improve. Machine Learning pp. 1–37 (2023) 2, 3
40. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.: Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering (2022) 1, 3

41. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer Vision–ECCV 2016: 14th European Conference. pp. 499–515. Springer (2016) 3
42. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017) 9
43. Xu, Q., Zhang, R., Fan, Z., Wang, Y., Wu, Y.Y., Zhang, Y.: Fourier-based augmentation with applications to domain generalization. Pattern Recognition **139**, 109474 (2023) 6
44. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021) 2, 3, 6
45. Yang, Y., Zha, K., Chen, Y., Wang, H., Katabi, D.: Delving into deep imbalanced regression. In: International Conference on Machine Learning. pp. 11842–11851. PMLR (2021) 3
46. Yao, H., Wang, Y., Zhang, L., Zou, J., Finn, C.: C-mixup: Improving generalization in regression. In: Proceeding of the Conference on Neural Information Processing Systems (2022) 2, 3, 6, 8, 9, 10
47. Zha, K., Cao, P., Son, J., Yang, Y., Katabi, D.: Rank-n-contrast: Learning continuous representations for regression. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 3
48. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018) 2, 3, 9
49. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2015) 1
50. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (2021) 3