





Multi-target Weakly Supervised Regression Using Manifold Regularization and Wasserstein Metric

Kirill Kalmutskiy^{1,2}(✉) , Lyailya Cherikbayeva³ , Alexander Litvinenko⁴ ,
and Vladimir Berikov^{1,2} 

¹ Sobolev Institute of mathematics, Novosibirsk, Russia
berikov@math.nsc.ru

² Novosibirsk State University, Novosibirsk, Russia
k.kalmutskii@ng.nsu.ru

³ Al-Farabi Kazakh National University, Almaty, Kazakhstan

⁴ RWTH Aachen, Aachen, Germany
litvinenko@uq.rwth-aachen.de
<http://www.uq.rwth-aachen.de>

Abstract. In this paper, we consider the weakly supervised multi-target regression problem where the observed data is partially or imprecisely labelled. The model of the multivariate normal distribution over the target vectors represents the uncertainty arising from the labelling process. The proposed solution is based on the combination of a manifold regularisation method, the use of the Wasserstein distance between multivariate distributions, and a cluster ensemble technique. The method uses a low-rank representation of the similarity matrix. An algorithm for constructing a co-association matrix with calculation of the optimal number of clusters in a partition is presented. To increase the stability and quality of the ensemble clustering, we use k-means with different distance metrics. The experimental part presents the results of numerical experiments with the proposed method on artificially generated data and real data sets. The results show the advantages of the proposed method over existing solutions.

Keywords: Weakly supervised learning · Multi-target regression · Manifold regularization · Low-rank matrix representation · Cluster ensemble · Co-association matrix

1 Introduction

Weakly supervised learning is a type of machine learning technique in which a model is trained using incomplete, imprecise, or ambiguous supervision signals, rather than using fully correctly labeled data. Weak supervision often arises in real problems for various reasons. This may be due to an expensive data labeling process, poor accuracy of sensors, insufficient expert qualifications or human error. For example, there is weak supervision in cases where the labeling

is obtained using crowdsourcing techniques: for each object there is a set of different (possibly inaccurate) labels, the quality of which depends on the skills of the performers. In addition to that, some objects may remain unlabeled if there is not enough budget for them.

Another example is the task of detecting objects in an image [1]. Bounding boxes are a common way to represent the location and extent of objects detected in an image or video frame in object detection tasks. A bounding box is a rectangular box that surrounds the object and is defined by its four corners or coordinates. In some difficult cases, such as detecting objects in medical CT scans, the bounding boxes can be very inaccurate and may highlight unwanted pixels. Moreover, the process of labeling CT images is very time-consuming, so it is not possible to label many objects.

Generally, there are three types of weak supervision: incomplete supervision, inaccurate supervision and inexact supervision [2]. In this work, we focus on the first two types of weak supervision. In particular, we assume that only a small part of the objects have labels, while the labels can be uncertain, and for most of the dataset there are no labels at all.

We propose an algorithm for solving the multi-target weakly supervised regression problem using Wasserstein metric, manifold regularization and a co-association matrix as the similarity matrix. We follow the transductive setting, which means that the objects from test data can be used during training and the task is to find the labels only for these objects. The algorithm for calculating the weighted average co-association matrix is also improved. Finally, we compare the proposed algorithm with existing algorithms of supervised learning and weakly supervised learning on synthetic and real data.

2 Problem Description

Let $X = \{x_1, \dots, x_n\}$, $x_i = (x_i^1, \dots, x_i^p)^\top \in \mathbb{R}^p$ are sampled from distribution \mathcal{P}_X , where n is the number of objects in the sample and p is the dimensionality of the feature space. In turn, $Y = \{y_1, \dots, y_n\}$, $y_i = (y_i^1, \dots, y_i^m)^\top \in \mathbb{R}^m$ are target labels, where m is the dimensionality of the target feature space.

In the semi-supervised transductive learning problem, a dataset $X \times Y = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is considered, but the target features $\{y_1, \dots, y_{n_1}\} = Y_1 \subseteq Y$ are only known for a small part of the available data $\{x_1, \dots, x_{n_1}\} = X_1 \subseteq X$. The rest of the objects $\{x_{n_1+1}, \dots, x_n\} = X_0 \subseteq X$ are unlabeled. The task is to predict the labels $Y_0 = \{y_{n_1+1}, \dots, y_n\}$ as accurately as possible according to some criterion.

To model the uncertainty of the observed labels, we use a multivariate normal distribution. We suppose that for each i -th data point, $i = 1, \dots, n_1$, the value y_i of the target feature is a realization of a random variable Y_i with a cumulative distribution function (cdf) $F_i(y)$ defined on $D_Y \subset \mathbb{R}^m$:

$$Y_i \sim \mathcal{N}(\mu_i, \Sigma_i), \quad (1)$$

where $\mu_i \in \mathbb{R}^m$ is a mean vector, $\Sigma_i \in \mathbb{R}^{m \times m}$ is a covariance matrix, $i = 1, \dots, n_1$. The overall degree of uncertainty can be interpreted as $\mathbb{T}_i = |\Sigma_i|$: the larger it is,

the greater the uncertainty of the label. Accordingly, for strictly labeled objects, it is expected that $\mathbb{T}_i \approx 0$.

The task is to determine $F_i(y)$ for $i = n_1 + 1, \dots, n$ following an objective criterion.

3 Related Work

The work [3] provides algorithms WSR-RBF and WSR-LRCM for solving the weakly supervised regression problem in the transductive formulation in the case of a one-dimensional target variable. It uses a univariate normal distribution to model inaccuracy:

$$Y_i \sim \mathcal{N}(a_i, \sigma_i),$$

where σ_i is an indicator of inaccuracy. Then it is proposed to solve the optimization problem by minimizing the distance between the predicted and real distributions using manifold regularization. To approximate the similarity matrix in WSR-LRCM, the co-association matrix is used and to obtain the co-association matrix, the cluster ensemble and the k-means algorithm are used. The WSR-RBF variant uses a weight matrix based on the RBF kernel instead of a low-rank representation:

$$W_{ij} := W(h) = \exp\left(-\frac{h^2}{2\ell^2}\right), \tag{2}$$

where $h = \|x_i - x_j\|$, and ℓ is a parameter.

However, the presented algorithm does not generalize to the multidimensional case. To solve a multi-target regression, it is necessary to train a separate model for each target variable. With this approach, it is possible to effectively solve those problems in which the target variables are independent of each other. If the target variables are not independent, for example, in the problem of object detection [1], these dependencies will be lost during training. These dependencies can be taken into account by using the distance between multivariate distributions, such as the Wasserstein distance [4].

The article [5] presents a detailed analysis of the co-association matrix and the algorithm for its construction. However, it relies on the basic version of the k-means algorithm, which has significant drawbacks, including the use of a single metric option and the uncertainty in choosing the appropriate number of clusters. In [7] the authors analyze the influence of metrics other than Euclidean on the quality of clustering by the k-means algorithm.

4 Proposed Method

Let

- $F^* = \{F_1^*, \dots, F_{n_1}^*, \dots, F_n^*\}$ be the set of arbitrary multivariate normal cdf's, each F_i^* is represented by a pair (a_i, \mathbb{S}_i) ;

– $F = \{F_1, \dots, F_{n_1}\}$ be the set of known cdf's, each F_i is represented by a pair (μ_i, Σ_i) .

In the following, we assume, both Σ_i and \mathbb{S}_i to be positive-definite matrices. Therefore, they are admitting Cholesky decomposition: $\Sigma_i = \Sigma_i^{1/2} \Sigma_i^{1/2\top}$, $\mathbb{S}_i = \mathbb{S}_i^{1/2} \mathbb{S}_i^{1/2\top}$. We denote elements of $\mathbb{S}_i^{1/2}$ as s_{jk}^i , and elements of $\Sigma_i^{1/2}$ as σ_{jk}^i .

4.1 Objective Functional

Consider the following optimization problem:

$$\text{find } F^{**} = \arg \min_{F^*} J(F, F^*), \text{ where}$$

$$J(F, F^*) = \sum_{x_i \in X_1} \mathcal{W}(F_i, F_i^*) + \gamma \sum_{x_i, x_j \in X} \mathcal{W}(F_i^*, F_j^*) W_{ij}$$

where \mathcal{W} is a 2-Wasserstein metric [4] (also known as Kantorovich-Rubenstein distance), $\gamma > 0$ is a parameter, and matrix $W = (W_{ij})$ represents the similarity measures between elements of dataset. For two multivariate Gaussian distributions $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$, 2-Wasserstein distance is

$$\mathcal{W}(N(\mu_0, \Sigma_0), N(\mu_1, \Sigma_1)) = \|\mu_0 - \mu_1\|_2^2 + \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_F^2.$$

Following [3], we also add the regularisation term with parameter $\beta > 0$. We can rewrite the objective as

$$\begin{aligned} \text{find } (a^*, \mathbb{S}^*) &= \arg \min_{(a, \mathbb{S})} J(\mu, \Sigma, a, \mathbb{S}), \text{ where} \\ J(\mu, \Sigma, a, \mathbb{S}) &= \sum_{x_i \in X_1} \|\mu_i - a_i\|_2^2 + \|\Sigma_i^{1/2} - \mathbb{S}_i^{1/2}\|_F^2 \\ &+ \gamma \sum_{x_i, x_j \in X} W_{ij} (\|a_i - a_j\|_2^2 + \|\mathbb{S}_i^{1/2} - \mathbb{S}_j^{1/2}\|_F^2) \\ &+ \beta \sum_{i=1, \dots, n} \|a_i\|_2^2 + \|\mathbb{S}_i\|_F^2. \end{aligned} \tag{3}$$

4.2 Optimal Solution

To find the optimal solution, we differentiate (3) with respect to elements of a_i and $\mathbb{S}_i^{1/2}$, $i = 1, \dots, n$:

$$\frac{\partial J}{\partial a_{ij}} = 2(\mu_{ij} - a_{ij}) + 4\gamma \sum_{l=1, \dots, n} W_{lj} (a_{lj} - a_{ij}) + 2\beta a_{ij}, \quad i = 1, \dots, n_1$$

$$\frac{\partial J}{\partial a_{ij}} = 4\gamma \sum_{l=1, \dots, n} W_{lj} (a_{lj} - a_{ij}) + 2\beta a_{ij}, \quad i = n_1, \dots, n$$

$$\frac{\partial J}{\partial s_{jk}^i} = 2(s_{jk}^i - \sigma_{jk}^i) + 4\gamma \sum_{l=1, \dots, n} W_{li} (s_{jk}^l - s_{jk}^i) + 2\beta s_{jk}^i, \quad i = 1, \dots, n_1$$

$$\frac{\partial J}{\partial s_{jk}^i} = 4\gamma \sum_{l=1, \dots, n} W_{li} (s_{jk}^l - s_{jk}^i) + 2\beta s_{jk}^i, \quad i = n_1, \dots, n.$$

Given that the matrices $\Sigma_i^{1/2}$ are lower triangular, we introduce an auxiliary operation $\text{vec}_2 : \mathbb{R}^{m \times m} \rightarrow \mathbb{R}^{\frac{m(m+1)}{2}}$ that transforms all elements above the main diagonal (including the main diagonal elements) into a row-by-row vector. Also, a lower triangular matrix can be obtained from a vector using an operation $\text{vec}_2^{-1} : \mathbb{R}^{\frac{m(m+1)}{2}} \rightarrow \mathbb{R}^{m \times m}$. Similarly, operation $\text{vec}_3 : \mathbb{R}^{n \times m \times m} \rightarrow \mathbb{R}^{n \times \frac{m(m+1)}{2}}$ (as well as $\text{vec}_3^{-1} : \mathbb{R}^{n \times \frac{m(m+1)}{2}} \rightarrow \mathbb{R}^{n \times m \times m}$) can be defined for three-dimensional tensors whose elements are lower triangular matrices. Let us denote

$$\begin{aligned} Y_{1,0} &= (\mu_1^\top, \dots, \mu_{n_1}^\top, 0, \dots, 0) \in \mathbb{R}^{n \times m} \\ \Sigma_{1,0} &= (\text{vec}_2(\Sigma_1^{1/2})^\top, \dots, \text{vec}_2(\Sigma_{n_1}^{1/2})^\top, 0, \dots, 0) \in \mathbb{R}^{n \times \frac{m(m+1)}{2}} \\ B &= \text{diag}(\beta + 1, \dots, \beta + 1, \beta, \dots, \beta) \in \mathbb{R}^{n \times n}. \end{aligned}$$

Then the solution of the optimization problem can be given in the matrix form

$$\begin{aligned} a^* &= (B + 2\gamma L)^{-1} Y_{1,0} \\ \mathbb{S}^* &= \text{vec}_3^{-1}((B + 2\gamma L)^{-1} \Sigma_{1,0}) \end{aligned} \tag{4}$$

where L is the Laplacian matrix, i.e., $L = D - W$, D is a diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$. If we assume that there is exist $V \in \mathbb{R}^{n \times q}$, $q \ll n$, such that $W = VV^\top$ then

$$B + 2\gamma L = B + 2\gamma D - 2\gamma VV^\top = G - 2\gamma VV^\top.$$

where $G = B + 2\gamma D$. By using the Woodbury identity [6], the inverse operator $B + 2\gamma L$ in the solution, that takes $O(n^3)$ operations, can be represented as

$$(G - 2\gamma VV^\top)^{-1} = G^{-1} + 2\gamma G^{-1}V(I - 2\gamma V^\top G^{-1}V)^{-1}V^\top G^{-1} \tag{5}$$

where G is diagonal matrix (and therefore can be inverted in linear time), $I - 2\gamma V^\top G^{-1}V \in \mathbb{R}^{q \times q}$. Therefore it takes $O(nq + q^3)$ to perform the inverse, which reduces the computations significantly, since by the assumption $q \ll n$. Finally, we get:

$$\begin{aligned} a^* &= (G^{-1} + 2\gamma G^{-1}V(I - 2\gamma V^\top G^{-1}V)^{-1}V^\top G^{-1})Y_{1,0} \\ \mathbb{S}^* &= \text{vec}_3^{-1}((G^{-1} + 2\gamma G^{-1}V(I - 2\gamma V^\top G^{-1}V)^{-1}V^\top G^{-1})\Sigma_{1,0}). \end{aligned} \tag{6}$$

In the article [3] it is shown that the weighted average co-association matrix can be used as a similarity matrix. By definition, the weighted average co-association matrix is

$$H = \sum_{l=1}^r \omega_l H_l, \tag{7}$$

where H_1, \dots, H_r are the co-association matrices for partitions P_1, \dots, P_r with elements indicating whether a pair x_i, x_j belong to the same cluster of this

partition or not, $\omega_1, \dots, \omega_r$ are weights of ensemble elements, $\omega_l \geq 0$, $\sum \omega_l = 1$. This matrix has a low-rank representation:

$$H = VV^\top,$$

where $V = [V_1 V_2 \dots V_r]$ is a block matrix, $V_l = \sqrt{\omega_l} Z_l$, $Z_l \in \mathbb{R}^{n \times K_l}$ is the cluster assignment matrix for l th partition: $Z_l(i, k) = \mathbb{I}[c(x_i) = k]$, $i = 1, \dots, n$, $k = 1, \dots, K_l$ and K_l is the number of clusters in partition P_l , $K_l \ll n$. It is also shown that the Laplacian matrix L for the matrix H can be written in the following form:

$$\begin{aligned} L &= D' - H, \\ D' &= \text{diag}(D'_{11}, \dots, D'_{nn}), \\ D'_{ii} &= \sum_{j=1}^n H(i, j) = \sum_{j=1}^n \sum_{l=1}^r \omega_l \sum_{k=1}^{K_l} Z_l(i, k) Z_l(j, k). \end{aligned} \quad (8)$$

Now the optimal solution (5) can be found by using the low-rank representation of the similarity matrix (7) and the diagonal matrix (8).

5 Co-association Matrix: Multimetricity and Optimality

To obtain a low-rank similarity matrix representation, we will use a weighted average co-association matrix as the similarity matrix. However, the standard algorithm for calculating the weighted average co-association matrix [5] has a number of disadvantages:

- The k-means algorithm using the Euclidean metric can only find spherical clusters, so some complex relationships in the data may not be found as a result of clustering;
- The result is strongly influenced by both the choice of the desired number of clusters for the k-means algorithm and the number of different partitions in the ensemble.

To solve these problems, we decided to improve the algorithm for calculating the weighted average co-association matrix. Firstly, we propose to average the co-association matrix over the distance metrics used in the k-means algorithm. Secondly, we propose to use only optimal partitions in terms of cluster validity index in the ensemble in order to reduce the influence of unnecessary partitions and reduce the size of the ensemble.

5.1 Multimetric Weighted Average Co-association Matrix

Let $\{M_t\}_{t=1}^d$ be the set of metrics that can be used in the k-means algorithm as the distance between points, for example, the Minkowski distance of order p . Then for each metric from this set, an arbitrary set of partitions variants $\{P_l^{M_t}\}_{l=1}^{r^{M_t}}$ can be obtained using cluster ensemble. Similarly, for each partition,

the co-association matrix $H_l^{M_t}$ can be found [3]. Then we define the multimetric weighted average co-association matrix as follows:

$$H = \sum_{t=1}^d H^{M_t} = \sum_{t=1}^d \sum_{l=1}^{r^{M_t}} \omega_l^{M_t} H_l^{M_t}, \quad (9)$$

where $\omega_1^{M_t}, \dots, \omega_r^{M_t}$ are weights of ensemble elements, $\omega_l^{M_t} \geq 0$, $\sum_{l=1}^{r^{M_t}} \omega_l^{M_t} = 1$ for each M_t , $t = 1, \dots, d$.

It should be noted that the clustering quality index, on which partition weights $\omega_l^{M_t}$ depend, should use the selected metric as the distance between points. That is why we assume that $\sum_{l=1}^{r^{M_t}} \omega_l^{M_t} = 1$ for each M_t , $t = 1, \dots, d$ rather than $\sum_{t=1}^d \sum_{l=1}^{r^{M_t}} \omega_l^{M_t} = 1$. As a further improvement, co-association matrices can also be weighted.

Thus, by using different metrics, we can obtain different partitions and reduce the impact of some negative effects arising from the use of the Euclidean distance. For example, in [7] it is shown that using the city blocks metric can reduce the impact of the curse of dimensionality.

5.2 Optimal Weighted Average Co-association Matrix

In general, the number of clusters in each partition is a hyperparameter. For example, in [3] two different set of parameters are used:

- The ensemble size $r = 10$, the number of clusters K_i in i -th partition: $K_i = 2 + i$, $i = 1, \dots, r$;
- The ensemble size $r = 10$, the number of clusters K_i in i -th partition: $K_i = 100 + i$, $i = 1, \dots, r$.

However, this choice may not be optimal. So, in the first case, for partitions with a small number of clusters, the weights can be extremely small, which means that their influence on the weighted average co-association matrix will be insignificant. In the second case, in addition to the high computational complexity of finding partitions with a large number of clusters, all resulting partitions can be similar to each other and have almost the same weights. Also, in both cases, it is not guaranteed that at least one optimal partition will be found in terms of any criterion: for example, a partition that achieves a local optimum of the cluster validity index.

We propose another algorithm that calculates weighted average co-association matrix with optimal partitions. The matrix H^* thus obtained is called optimal weighted average co-association matrix. This matrix is optimal in the sense that only optimal partitions according to the cluster validity index are used in its calculation. Below is an algorithm for calculating the optimal weighted average co-association matrix by steps:

Input: \mathbf{X} - dataset. \mathbf{r} - cluster ensemble size. k_{\min} - minimum number of clusters in a partition. k_{\max} - maximum number of clusters in a partition.**Output:** H^* - optimal weighted average co-association matrix.**Steps:**

1. Find a set of partitions $\{P_k\}_{k=k_{\min}}^{k_{\max}}$ of \mathbf{X} using the k-means algorithm with different number of clusters k .

2. Calculate a set of cluster validity index values $\{\omega_k\}_{k=k_{\min}}^{k_{\max}}$ for the set of partitions $\{P_k\}_{k=k_{\min}}^{k_{\max}}$.

3. Select \mathbf{r} largest values $\{\omega_{k_i}\}_{i=1}^r$ from a set $\{\omega_k\}_{k=k_{\min}}^{k_{\max}}$ and the corresponding set of partitions $\{P_{k_i}\}_{i=1}^r$.

4. Calculate a set of co-association matrices $\{H_{k_i}\}_{i=1}^r$ for the set of partitions $\{P_{k_i}\}_{i=1}^r$.

5. Calculate optimal weighted average co-association matrix $H^* = \sum_{l=1}^r \omega_{k_l} H_{k_l}$

end.

The optimal weighted average co-association matrix thus obtained can be used instead of the original one, including for calculating multimetric weighted average co-association matrix:

$$H^* = \sum_{t=1}^d H^* M_t. \quad (10)$$

6 C-WSR Algorithm

We formulate three main variants of the Correlated Weakly Supervised Regression (C-WSR) algorithm:

- **RBF**: Radial Basis Function to calculate the similarity matrix is used;
- **LRCM**: a low-rank representation of the weighted average co-association matrix to calculate the similarity matrix is used;
- **LROMCM**: a low-rank representation of the optimal multimetric weighted average co-association matrix (10) to calculate the similarity matrix is used.

Input:

\mathbf{X} - dataset with weak supervision, $X_1 \subset \mathbf{X}$ - labeled sample, $X_2 \subset \mathbf{X}$ inaccurately labeled sample, $X_3 \subset \mathbf{X}$ - unlabeled sample.

a_i, Σ_i - mean vectors and covariance matrices of target distributions for each $x_i \in X_1 \cup X_2$

LRCM variant: r, Ω - cluster ensemble size and set of parameters for the k-means for clustering.

LRMCM variant: M - set of metrics for algorithm k-means, r - cluster ensemble size, k_{\min} - minimum number of clusters in a partition, k_{\max} - maximum number of clusters in a partition.

Output:

a^* , \mathbb{S}^* - predicted mean vectors and covariance matrices of target distributions for objects from sample \mathbf{X} (including predictions for the unlabeled sample).

RBF Variant Steps:

Directly calculate predicted mean vectors and covariance matrices of target distributions using (2) and (4).

LRCM Variant Steps:

1. Generate r variants of clustering partition for parameters randomly chosen from Ω ; calculate weighted average co-association matrix.
2. Find graph Laplacian in the low-rank representation using (7) and D' in (8).
3. Calculate predicted mean vectors and covariance matrices of target distributions using (6).

LRMCM Variant Steps:

1. Calculate optimal multimetric weighted average co-association matrix with metrics from set M and parameters r , k_{\min} , k_{\max} using (9) and (10).
2. Find graph Laplacian in the low-rank representation using (7) and D' in (8).
3. Calculate predicted mean vectors and covariance matrices of target distributions using (6).

end.

7 Experimental Results

In this section, we will compare three variants of the proposed Correlated Weakly Supervised Regression (C-WSR) algorithm. We use the MWD metric when comparing with weakly supervised learning algorithms WSR-RBF and WSR-LRCM from [3] and MAE when comparing with supervised learning algorithms such as Multivariate Linear Regression and gradient boosting from framework XGBoost on real data:

$$\text{MWD}(y, y^*) = \frac{1}{n_{test}} \sum_{x_i \in X_{test}} \|\mu_i - a_i\|_2^2 + \|\Sigma_i^{1/2} - \mathbb{S}_i^{1/2}\|_F^2,$$

$$\text{MAE}(y, y^*) = \frac{1}{n_{test}} \sum_{x_i \in X_{test}} \|\mu_i - a_i\|_2.$$

Since the WSR-RBF and WSR-LRCM algorithms can only be used in a single target scenario, we train a separate model for each target variable. To calculate multimetric weighted average co-association matrix, we use Minkowski metric ρ_p with different $p \in \{1, 2, \infty\}$ and Silhouette as index cluster validity to determine the weights and the optimal number of clusters.

For experiments, we used an AMD Ryzen 9 3850X processor with a clock frequency of 3.5 GHz and 64 GB of RAM.

7.1 Monte-Carlo Simulation

For the Monte Carlo simulation, we generated a dataset of 1000 objects from a mixture of multivariate normal distributions $\mathcal{N}(\mu_k^*, \Sigma_k^*)$, $\mu_k^* = (8k + 1, 8k + 2, \dots, 8k + d_x) \in \mathbb{R}^m$, $\Sigma_k^* = \text{diag}(1, \dots, 1) \in \mathbb{R}^{d_x \times d_x}$, $d_x = 8$ and $k \in \{1, 2, 3\}$.

For objects generated from the k -th component, we assume that the target function is equal to $Y_k = k + \varepsilon_k$, where ε_k is a random variable with d_y -dimensional normal distribution function $\mathcal{N}(0, D_k D_k^\top)$, D_k is random lower-triangular matrix with elements sampled from normal distribution and $d_y = 4$.

To insure the weak supervision, we assumed 10% of the dataset to be strictly labeled, 20% of the dataset consists of inaccurately labeled objects and the remaining 70% of objects are unlabeled. To model the inaccurate labeling, we use the parameters defined in (1): $\Sigma_i = \Sigma_Y$, where Σ_Y is a covariance matrix of the target function over labeled data. For strictly labeled objects, we assume that the matrix Σ_i is a zero matrix.

For the WSR-LRCM and C-WSR-LRCM algorithms, we used a cluster ensemble of size $r = 30$ and the number of clusters K_i in i -th partition: $K_i = 2 + i$, $i = 1, \dots, 30$. The C-WSR-LROMCM algorithm uses parameters $r = 10$, $k_{\min} = 2$ and $k_{\max} = 30$. Regularization coefficients $\beta = 0.001$ and $\gamma = 0.001$ are set for all algorithms. The obtained quality metrics were averaged over 100 runs. The results are presented in Table 1.

Table 1. Comparison on Monte-Carlo simulation.

Supervision type	WSR		C-WSR		
	RBF	LRCM	RBF	LRCM	LROMCM
MWD	0.835	0.760	0.382	0.324	0.227

7.2 CO/NOx Dataset

For CO/NOx dataset [8] we use carbon monoxide (CO) and nitrogen oxides (NOx) emissions for year 2015 as regression targets. This dataset contains 11 features that describe the characteristics of a gas turbine and include 36733 observations.

1% of data is assumed to be strictly labeled, 9% is assumed to be labeled inaccurately, and 90% of data is considered unlabeled. Since the dataset is large, to model the inaccurate labeling, we estimate the mean vectors μ_i and covariance matrices Σ_i by 50 nearest neighbours. For strictly labeled objects, the exact label is used as the mean vector μ_i , and Σ_i is equal to the zero matrix.

As with synthetic data, regularization coefficients $\beta = 0.001$ and $\gamma = 0.001$ are set. For the WSR-LRCM and C-WSR-LRCM algorithms, a cluster ensemble of size $r = 30$ is used with the number of clusters K_i in i -th partition: $K_i = 10 + i$, $i = 1, \dots, 30$. The C-WSR-LROMCM algorithm trained with parameters $r = 10$,

$k_{\min} = 2$ and $k_{\max} = 50$. Supervised learning algorithms (Multivariate Linear Regression (MLR) and XGBoost (XGB)) are trained only on strictly labeled objects. Note that due to the large amount of data in the dataset, finding the inverse matrix in the RBF variant requires a significant amount of computing resources, especially RAM. The results are presented in Table 2.

Table 2. Comparison on CO/NOx dataset.

Supervision type	WSR		C-WSR			SR	
	RBF	LRCM	RBF	LRCM	LROMCM	MLR	XGB
MWD	72.96	60.07	65.45	52.22	44.74	–	–
MAE	42.11	35.92	38.84	31.92	26.83	38.69	30.48

Thus the results of the experiments show the considerable improvements in the accuracy for the proposed method.

8 Conclusion

In this paper, we considered the problem of multi-target weakly supervised regression with noisy labelling in a transductive setting. Using the multivariate normal distribution, we described an imprecision model in the multi-output case. We also proposed an algorithm for solving the optimisation problem using the Wasserstein metric and manifold regularisation. To speed up the solution of the optimisation problem, we used the cluster ensemble to obtain the co-association matrix and the low-rank representation technique to compress the resulting matrices.

The presented algorithm has shown its advantage over existing machine learning algorithms that cannot use uncertain multidimensional labels during training. We have also made several important improvements to the calculation of the weighted average co-association matrix by introducing an optimal multivariate weighted average co-association matrix. The new approach can significantly improve the quality and stability of the algorithm, and also simplifies the search for optimal hyperparameters to solve each specific problem.

As a further improvement, one can try different distances between distributions in the optimisation problem, and another promising idea would be to use deep learning approaches to find the co-association matrix. It is also worth considering other imprecision models: for example, using different types of multivariate distributions than normal.

Acknowledgements. The work was carried out with the financial support of the Russian Science Foundation, project 22-21-00261. Special thanks to Vladimir Kondratiev for participating in the discussion and experiments.

References

1. Yang, Z., Mahajan, D., Ghadiyaram, D., Nevatia, R., Ramanathan, V.: Activity driven weakly supervised object detection, pp. 2912–2921 (2019). <https://doi.org/10.1109/CVPR.2019.00303>
2. Zhou, Z.H.: A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**, 44–53 (2017)
3. Berikov, V., Litvinenko, A.: Weakly supervised regression using manifold regularization and low-rank matrix representation. In: Pardalos, P., Khachay, M., Kazakov, A. (eds.) *MOTOR 2021*. LNCS, vol. 12755, pp. 447–461. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77876-7_30
4. Bogachev, V.I., Kolesnikov, A.: The Monge-Kantorovich problem: achievements, connections, and perspectives *Russ. Math. Surv.* **67**, 785–890 (2012)
5. Berikov, V.B.: Cluster ensemble with averaged co-association matrix maximizing the expected margin. In: *International Conference on Discrete Optimization and Operations Research (DOOR 2016)*, vol. 1623, pp. 489–500. CEUR-WS.org (2016)
6. Higham N.: *Accuracy and Stability of Numerical Algorithms*. SIAM (2002)
7. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) *ICDT 2001*. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_27
8. Kaya, H., Tüfekci, P., Üzun, E.: Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS. *Turk. J. Electr. Eng. Comput. Sci.* **27**, 4783–4796 (2019)