Sinkhorn Regression

Lei Luo[1,2] , **Jian Pei**[3] , **Heng Huang**[1,2*]

[1] JD Finance America Corporation, Mountain View, CA, USA
[2]Department of Electrical and Computer Engineering, University of Pittsburgh, PA, USA
[3]School of Computing Science, Simon Fraser University, Canada
lei.luo@jd.com, jpei@cs.sfu.ca, heng.huang@pitt.edu

Abstract

This paper introduces a novel Robust Regression (RR) model, named Sinkhorn regression, which imposes Sinkhorn distances on both loss function and regularization. Traditional RR methods target at searching for an element-wise loss function (*e.g.*, L_p-norm) to characterize the errors such that outlying data have a relatively smaller influence on the regression estimator. Due to the neglect of the geometric information, they often lead to the sub-optimal results in the practical applications. To address this problem, we use a cross-bin distance function, *i.e.*, Sinkhorn distances, to capture the geometric knowledge from real data. Sinkhorn distances is invariant in movement, rotation and zoom. Thus, our method is more robust to variations of data than traditional regression models. Meanwhile, we leverage Kullback-Leibler divergence to relax the proposed model with marginal constraints into its unbalanced formulation to adapt more types of features. In addition, we propose an efficient algorithm to solve the relaxed model and establish its complete statistical guarantees under mild conditions. Experiments on the five publicly available microarray data sets and one mass spectrometry data set demonstrate the effectiveness and robustness of our method.

1 Introduction

Regression analysis is an important statistical technique frequently applied in machine learning for a large variety of tasks such as image classification [Naseem *et al.*, 2010] and subspace segmentation [Lu *et al.*, 2012]. However, traditional linear regression models (*e.g.*, least square regression) are mainly based on the normal population, which cannot provide the robustness against outlier. To address this problem, robust regression has emerged as a powerful tool to handle the data with non-Gaussian noises. The representative models include Robust Sparse Coding (RSC) [Yang *et al.*, 2011] and Correntropy based Sparse Representation (CESR) [He *et al.*, 2011]. They can be viewed as M-estimation problems induced by different influence functions which can make outlying data have a relatively smaller influence on the regression estimator.

However, those robust regression methods ignore the spatial structure information of real noises since they usually assume that noise pixels are independently generated. To alleviate this problem, [Li *et al.*, 2013] proposed the Structured Sparse Error Coding (SSEC) model by exploring the intrinsic structure of continuous occlusion. [Jia *et al.*, 2012] used a class of structured sparsity inducing norms to fit structural noises. [Yang *et al.*, 2017] presented the two-dimensional image-matrix-based error models by employing nuclear norm to characterize the practical noise matrix. Although these methods have shown great potential in handling structural noises such as occlusion and illumination, they cannot characterize the geometric knowledge that often exists in real data and is helpful for improving the model performance.

More recently, Wasserstein distance, derived from the Optimal Transport (OT) theory, has drawn ample attention in many machine learning tasks. Differing from L_p distances ($p \geq 1$) or Kullback-Leibler and other f-divergences, which require distributions to be absolutely continuous with respect to each other or to a base measure, Wasserstein distance is well-defined between any pair of probability distributions over a sample space equipped with a metric. Thus, it provides a meaningful notion of closeness (*i.e.,* distance) for distributions supported on non-overlapping low dimensional manifolds. Due to this advantage, Wasserstein distance has been successfully applied to cancer detection [Ozolek *et al.*, 2014] and super-resolution [Kolouri and Rohde, 2015] problems. In addition, Kusner et al. [Kusner *et al.*, 2015] proposed the Word Mover's Distance (WMD), an implementation of the Wasserstein distance for textual data. To capture both global (at distribution scale) and local (at samples' scale) interactions between classes, Flamary et al. [Flamary *et al.*, 2016] described a new Wasserstein Discriminant Analysis by using a mechanism that can induce neighborhood preservation.

Motivated by the success of Wasserstein distances, in this paper, we propose a novel efficient and robust Matrix Regression method by employing joint Sinkhorn distances (*i.e.*, Wasserstein distances with Entropic constraints) minimization on both loss function and regularization. This is the first

*Corresponding Author. H. Huang was partially supported by U.S. NSF IIS 1836945, IIS 1836938, IIS 1845666, IIS 1852606, IIS 1838627, IIS 1837956. J. Pei was partially supported by NSERC Discovery Grants Program.

time for exploiting Sinkhorn distances to characterize both error and coefficient matrix in the regression model. To provide the high flexibility for non-distribution features (*e.g.*, original data features), we relax the proposed new model into its unbalanced formulation by virtue of Kullback-Leibler divergence. According to the OT geometry [Janati *et al.*, 2018], we know that using Sinkhorn distance as regularizer can promote parameters that are close since it takes into account a prior geometric knowledge on the regressor variables. As a result, our method is more reliable and applicable than traditional regression method using L_p-norm regularizer. To solve SMR, we derive an efficient algorithm based on Sinkhorn iteration, which iterates through applications of alternating optimization. Moreover, as the theoretical contribution of this paper, we provide a statistical bound for the proposed new model by leveraging the Rademacher complexity. We apply the proposed algorithm to feature selection problem, and make comparisons with some existing approaches. The experimental results show that the proposed method is more effective than the state-of-the-art methods.

2 Preliminaries

We summarize the notations and definitions used in this paper. Matrices are written as boldface uppercase letters. Vectors are written as boldface lowercase letters. $\| \cdot \|_F$ and $\| \cdot \|_*$ denote Frobenius norm and nuclear norm [Liu *et al.*, 2010], respectively. $\langle \cdot, \cdot \rangle$ is the inner product operation. $\mathbf{e}_{N_S} \in \mathcal{R}^{N_S}$ is a column vector of ones. $\mathcal{R}_+$ denotes the set of all non-negative real number. $\mathcal{R}_+^{N_S \times N_G}$ denotes the set of all positive semi-define matrices on $\mathcal{R}^{N_S \times N_G}$. For matrix $\mathbf{M} \in R^{n \times m}$, its *i*-th row, *j*-th column are denoted by $\mathbf{m}^i$, $\mathbf{m}_j$, respectively. The $L_{2,1}$-norm of $\mathbf{M}$ is defined as $\| \mathbf{M} \|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \| \mathbf{m}^i \|_2$, where $\| \cdot \|_2$ denotes the L_2-norm. We define the Kullback-Leibler (KL) divergence [Janati *et al.*, 2018] between two positive vectors $\mathbf{x}$ and $\mathbf{y}$ by

$$KL(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y}) \rangle + \langle \mathbf{y} - \mathbf{x}, \mathbf{e} \rangle. \quad (1)$$

The KL divergence between two matrices $\mathbf{A} \in \mathcal{R}_+^{N_S \times N_S}$ (*i.e.*, $a_{ij} \geq 0$ for all (i,j)) and $\mathbf{B} \in \mathcal{R}_{++}^{N_S \times N_S}$ (*i.e.*, $b_{ij} > 0$ for all (i,j)) is defined as

$$KL(\mathbf{A}|\mathbf{B}) = \sum_{i,j}^{N_S} a_{ij}(\log(\frac{a_{ij}}{b_{ij}}) - 1). \quad (2)$$

Optimal transport: OT theory, originally used to study the problem of resource allocation, provides a powerful geometrical tool for comparing probability distributions.

In a more formal way, given access to two sets of points $\mathcal{X}_S = \{\mathbf{x}_i^S \in \mathcal{R}^d\}_{i=1}^{N_S}$ and $\mathcal{X}_G = \{\mathbf{x}_i^G \in \mathcal{R}^d\}_{i=1}^{N_G}$, we construct two discrete empirical probability distributions as:

$$\hat{\mu}_S = \sum_{i=1}^{N_S} p_i^S \delta_{\mathbf{x}_i^S} \text{ and } \hat{\mu}_G = \sum_{i=1}^{N_G} p_i^G \delta_{\mathbf{x}_i^J}, \quad (3)$$

where p_i^S and p_i^G are probabilities associated to $\mathbf{x}_i^S$ and $\mathbf{x}_i^G$ respectively and $\delta_{\mathbf{x}}$ is a Dirac measure that can be interpreted as an indicator function taking value 1 at the position of $\mathbf{x}$ and 0 elsewhere. For these two distributions, the optimal transport (or Monge-Kantorovich) problem consists in finding a probabilistic coupling defined as a joint probability measure over $\mathcal{X}_S \times \mathcal{X}_G$ with marginals $\hat{\mu}_S$ and $\hat{\mu}_G$ that minimizes the cost of transport with respect to some *ground metric* $D : \mathcal{X}_s \times \mathcal{X}_g \to \mathcal{R}^+$:

$$\min_{\mathbf{P} \in \mathcal{U}_{\hat{\mu}_S, \hat{\mu}_G}} < \mathbf{P}, \mathbf{C} >, \quad (4)$$

where $\mathcal{U}_{\hat{\mu}_S, \hat{\mu}_G} = \{\mathbf{P} \in \mathcal{R}_+^{N_S \times N_G} : \mathbf{P}\mathbf{e}_{N_G} = \mathbf{p}^S, \ \mathbf{P}^T\mathbf{e}_{N_S} = \mathbf{p}^G\}$ is a set of doubly stochastic matrices, $\mathbf{p}^S = [p_1^S, p_2^S, \cdots, p_{N_S}^S]^T$ and $\mathbf{p}^G = [p_1^G, p_2^G, \cdots, p_{N_G}^G]^T$. $\mathbf{C} \in \mathcal{R}_+^{N_S \times N_G}$ is a given cost matrix, *i.e.*, $c_{i,j} = D(\mathbf{x}_i^S, \mathbf{x}_j^G)$, defining the energy needed to move a probability mass from $\mathbf{x}_i^S$ to $\mathbf{x}_j^G$. This problem admits a unique solution $\mathbf{P}^*$ and defines a metric on the space of probability measures (called the *Wasserstein distance*) as

$$W(\hat{\mu}_S, \hat{\mu}_G) = \min_{\mathbf{P} \in \mathcal{U}_{\hat{\mu}_S, \hat{\mu}_G}} < \mathbf{P}, \mathbf{C} > . \quad (5)$$

Entropic regularization: Solving problems with Wasserstein distance fitting errors can require solving several costly optimal transport problems. As a minimum of affine functions, the Wasserstein distance itself is not a smooth function. To avoid both of these issues, [Cuturi, 2013] proposed to smooth the optimal transport problem with an entropic term:

$$W_\gamma(\hat{\mu}_S, \hat{\mu}_G) = \min_{\mathbf{P} \in \mathcal{U}_{\hat{\mu}_S, \hat{\mu}_G}} < \mathbf{P}, \mathbf{C} > -\gamma e(\mathbf{P}). \quad (6)$$

where $\gamma > 0$ and e is the (strictly concave) entropy function:

$$e(\mathbf{P}) = -\langle \mathbf{P}, \log \mathbf{P} \rangle. \quad (7)$$

(6) allows to solve the optimal transportation problem efficiently using Sinkhorn-Knopp matrix scaling algorithm [Cuturi, 2013]. Thus, (6) is also called Sinkhorn distance, which preserve all advantages of Wasserstein distance. Compared to the original case (5), it can obtain smoother and more numerically stable solutions.

Matrix regression: In data mining and machine learning, a common paradigm for matrix regression is to minimize the penalized empirical loss:

$$\min_{\mathbf{Z}} L(\mathbf{Y} - \mathbf{A}^T\mathbf{Z}) + \lambda \Omega(\mathbf{Z}), \quad (8)$$

where $\lambda > 0$ is the balance parameter, $\mathbf{Z} \in \mathcal{R}^{m \times n}$ is the parameter to be estimated from the training sample matrix $\mathbf{A} \in \mathcal{R}^{m \times l}$ and the response matrix $\mathbf{Y} \in \mathcal{R}^{l \times n}$, $L(\mathbf{Y} - \mathbf{A}\mathbf{Z})$ is the empirical loss on the training set, and $\Omega(\mathbf{Z})$ is the regularization term that encodes feature relatedness. Different assumptions on the loss $\mathbf{Y} - \mathbf{A}\mathbf{Z}$ and variate $\mathbf{Z}$ can lead to different models. For example, if both $L(\cdot)$ and $\Omega(\cdot)$ are $L_{2,1}$-norm, (6) becomes the feature selection model based on $L_{2,1}$-norms [Nie *et al.*, 2010]; if $L(\cdot)$ is L_1-norm [Yang *et al.*, 2012; Yang *et al.*, 2013] and $\Omega(\cdot)$ is nuclear norm, (6) leads to the Low Rank Representation (LRR) model [Liu *et al.*, 2010]; if both $L(\cdot)$ and $\Omega(\cdot)$ are the square of Frobenius norm, (6) is the well-known Least Squares Regression (LSR) model. These models have been widely applied to many tasks such as feature selection, subspace segmentation and image classification.

3 Matrix Regression Based on Joint Sinkhorn Distances

In the above formulations, the loss term and estimated variate are characterized via the simple matrix norm. Thus, these models can be easily solved by conventional convex optimization methods (*e.g.*, ADMM [Liu *et al.*, 2010] and reweighted iterative methods [Nie *et al.*, 2010]). However, they suffer from two limitations. First, these matrix norm cannot offer the exibility in adapting them to various data types since they are nonparametric. Secondly, they do not taking into account the geometry of the data through the pairwise distances between the distributions' points. Accordingly, these models often achieve the suboptimal results in practical applications, especially when real data is corrupted by noise.

Sinkhorn Regression. Comparing with matrix norm, Sinkhorn distance can circumvent the above limitations (as stated in the introduction). Therefore, we propose to use Sinkhorn distance to jointly characterize loss term and estimated variate $\mathbf{Z}$, which is formulated as

$$\min_{\mathbf{Z},\mathbf{d}} \sum_{i=1}^{l} W_\gamma(h(\mathbf{Z}^T\mathbf{a}_i), h(\mathbf{y}^{i^T})) + \lambda \sum_{i=1}^{m} W_\gamma(h(\mathbf{z}^{i^T}), \mathbf{d}), \tag{9}$$

where $h(\cdot)$ denotes the histogram operator and the latent variable $\mathbf{d}$ consists in estimating the barycenter of $\{h(\mathbf{z}^{1^T}), h(\mathbf{z}^{2^T}), \cdots, h(\mathbf{z}^{m^T})\}$.

In model (9), the histogram operator is used to constrain $\mathbf{Z}^T\mathbf{a}_i$, $\mathbf{Y}^T$ and $\mathbf{Z}^T$, respectively. Thus, we do not need to extract the histogram features of data in the preprocessing stage. Meanwhile, since histogram features are invariant in movement, rotation and zoom, our method is more robust to the real noisy data than the other methods using matrix norm over original data.

Remark 1. Sinkhorn distance is used to compare row vectors of $\mathbf{A}^T\mathbf{Z}$ and $\mathbf{Y}$ in model (9). In fact, Sinkhorn distance can also be imposed on the corresponding column vectors, which induces the following model:

$$\min_{\mathbf{Z},\mathbf{d}} \sum_{i=1}^{n} W_\gamma(h(\mathbf{A}^T\mathbf{z}_i), h(\mathbf{y}_i)) + \lambda \sum_{i=1}^{n} W_\gamma(h(\mathbf{z}_i), \mathbf{d}).$$

The algorithm for solving the above model is similar to Algorithm 1. Due to the space limitations, we only focus on model (9) in this paper.

The proposed algorithm. Solving problem (9) is extremely challenging since it not only contains the composition of $h(\cdot)$ and $W(\cdot,\cdot)$, but also the computations of Sinkhorn distances with regard to different terms. Some existing algorithms [Rolet *et al.*, 2016; Sommerfeld *et al.*, 2018] are only suitable for solving Wasserstein loss minimization with matrix norm regularizer (*e.g.*, L_1- and $L_{2,1}$-norm). To cope with this challenge, we relax the marginal constraints $\mathcal{U}_{\hat{\mu}_S,\hat{\mu}_G}$ in (6) using a Kullback-Leibler divergence from the matrix to target marginals $\hat{\mu}_S$ and $\hat{\mu}_G$ is [Frogner *et al.*, 2015],

Algorithm 1 Solving (13) via Alternating Optimization

Input: data matrix $\mathbf{A}$ and $\mathbf{Y}$, parameters λ, γ, μ and ρ
Initialization: $\mathbf{P}^0$ and $\hat{\mathbf{P}}^0$
Output: model parameter $\mathbf{Z}$
Repeat
 for $i = 1$ to m **do**
  Update each $\mathbf{z}^i$ with proximal coordinate descent.
 end for
 for $i = 1$ to l **do**
  Update each $\mathbf{u}^i$ with proximal coordinate descent.
 end for
 Update $\mathbf{d}$ with proximal coordinate descent.
 Update $\mathbf{P}_1, \cdots, \mathbf{P}_l, \hat{\mathbf{P}}_1, \cdots, \hat{\mathbf{P}}_m$ via Sinkhorn iteration (Algorithms 2)
until convergence

i.e., (5) is converted as

$$W_\gamma(\hat{\mu}_S, \hat{\mu}_G) = \min_{\mathbf{P}\in\mathcal{R}_+^{N_S\times N_G}} \gamma KL(\mathbf{P}|\mathbf{K}) + \mu KL(\mathbf{P}\mathbf{e}_{N_G}|\hat{\mu}_S) + \mu KL(\mathbf{P}^T\mathbf{e}_{N_S}|\hat{\mu}_G), \tag{10}$$

where $\mathbf{K} = \exp(-\mathbf{C}/\gamma)$ and $\mathbf{C}$ is defined in Eq.(4). Let

$$f_{\hat{\mu}_S,\hat{\mu}_G}(\mathbf{P}) = \gamma KL(\mathbf{P}|\mathbf{K}) + \mu KL(\mathbf{P}\mathbf{e}_{N_G}|\hat{\mu}_S) + \mu KL(\mathbf{P}^T\mathbf{e}_{N_S}|\hat{\mu}_G), \tag{11}$$

and assume $\mathbf{A}, \mathbf{Y} > 0$. Then model (9) ultimately becomes:

$$\min_{\mathbf{Z},\mathbf{d},\mathbf{P}_1,\cdots,\mathbf{P}_l,\hat{\mathbf{P}}_1,\cdots,\hat{\mathbf{P}}_m} \sum_{i=1}^{l} f_{\mathbf{Z}\mathbf{a}_i,\mathbf{y}^{i^T}}(\mathbf{P}_i) + \lambda \sum_{i=1}^{m} f_{\mathbf{z}^{i^T},\mathbf{d}}(\hat{\mathbf{P}}_i). \tag{12}$$

where each $\mathbf{z}^i > 0$, $i = 1, 2, \cdots, m$.

To facilitate the design of the algorithm, we bring into an auxiliary variable $\mathbf{U} = \mathbf{A}^T\mathbf{Z}$ and rewrite model (12) as:

$$\min_{\mathbf{U},\mathbf{Z},\mathbf{d},\mathbf{P}_1,\cdots,\mathbf{P}_l,\hat{\mathbf{P}}_1,\cdots,\hat{\mathbf{P}}_m} \sum_{i=1}^{l} f_{\mathbf{u}^{i^T},\mathbf{y}^{i^T}}(\mathbf{P}_i) + \lambda \sum_{i=1}^{m} f_{\mathbf{z}^{i^T},\mathbf{d}}(\hat{\mathbf{P}}_i) + \rho \parallel \mathbf{A}^T\mathbf{Z} - \mathbf{U} \parallel_F^2 . \tag{13}$$

where $\rho > 0$.

The block coordinate descent method can be used to solve (13). This method alternates the minimization with respect to the variables $\{\mathbf{P}_1, \cdots, \mathbf{P}_l, \hat{\mathbf{P}}_1, \cdots, \hat{\mathbf{P}}_m\}$, $\mathbf{Z}$, $\mathbf{U}$ and $\mathbf{d}$. It is easy see that these variables are independent. Thus, they can be updated in parallel. The main iterative procedure for solving model (13) is summarized in Algorithm 1. In the following, we give the details for optimizing each variable.

Updating coefficient matrix $\mathbf{Z}$. For each $\mathbf{z}^i$, fixing other variables, problem (13) is simplified as:

$$\min_{\mathbf{z}_i} \lambda\mu KL(\hat{\mathbf{P}}_i\mathbf{e}|\mathbf{z}^{i^T}) + \rho \parallel \mathbf{A}^T\mathbf{Z} - \mathbf{U} \parallel_F^2 . \tag{14}$$

Considering the definition of KL divergence, (14) becomes

$$\min_{\mathbf{z}_i} \lambda\mu(\mathbf{z}^{i^T} - \hat{\mathbf{P}}_i\mathbf{e}\log\mathbf{z}^{i^T}) + \rho \parallel \mathbf{A}^T\mathbf{Z} - \mathbf{U} \parallel_F^2 . \tag{15}$$

Algorithm 2 Sinkhorn iteration for optimizing $\mathbf{P}_1, \cdots, \mathbf{P}_l$

Input: data matrix $\mathbf{A}$, coefficient matrix $\mathbf{Z}$ and auxiliary variable $\mathbf{U}$
Output: $\mathbf{P}_1, \cdots, \mathbf{P}_l$
for $i = 1$ to l **do**
$\mathbf{K}_i = \exp(-\mathbf{C}_i/\gamma)$, where $\mathbf{C}_i$ is the ground metric matrix between $\mathbf{u}^{i^T}$ and $\mathbf{y}^{i^T}$.
repeat
$\mathbf{w}_i \leftarrow \mathbf{u}^{i^T}/\mathbf{K}\mathbf{v}_i$; $\mathbf{v}_i \leftarrow \mathbf{y}^{i^T}/\mathbf{K}^T\mathbf{w}_i$.
until convergence
$\mathbf{P}_i \leftarrow (p_{(i)jt})_{n\times n}$, where the (j,t)-th element of $\mathbf{P}_i$ is $p_{(i)jt} = w_{i(j)}k_{(i)jt}v_{i(t)}$ ($w_{i(j)}$ is the *j*-th element of $\mathbf{w}_i$).
end for

To solve problem (15), we introduce the following Lemma:

Lemma 1. [Janati *et al.*, 2018] Let $a, b, \alpha \in \mathcal{R}_+$. The function $\eta : x \to (x - a\log(x)) + bx$ is convex on $\mathcal{R}_+$ and its proximal operator can be obtained in closed form, *i.e.*,

$$\text{prox}_{\alpha\eta(\cdot)}(y) = \frac{1}{2}[-\alpha(b+1)+y+\sqrt{(\alpha(b+1)-y)^2+4\alpha a}]. \tag{16}$$

Let

$$\varsigma(\mathbf{z}_i) = \lambda\mu(\mathbf{z}^{i^T} - \hat{\mathbf{P}}_i\mathbf{e}\log\mathbf{z}^{i^T}). \tag{17}$$

By Lemma 1, we can easily obtain the proximal operator $\text{prox}_{\varsigma(\cdot)}$. Therefore, using proximal coordinate descent, each $\mathbf{z}^{i^T}$ can be updated by

$$\mathbf{z}^{i^T} \leftarrow \text{prox}_{\varsigma(\cdot)}(\mathbf{z}^{i^T} - \delta(\mathbf{Z}^T(\mathbf{A}\mathbf{A}^T)_{(i)} - \mathbf{y}^{i^T})), \tag{18}$$

where $\delta > 0$ is the step size and $(\mathbf{A}\mathbf{A}^T)_{(i)}$ is the i-column of $\mathbf{A}\mathbf{A}^T$.

*Updating auxiliary variable **U**.* For each iteration, we obtain the optimal $\mathbf{u}^i$ by solving

$$\min_{\mathbf{u}^i} \mu KL(\mathbf{P}_i\mathbf{e}|\mathbf{u}^{i^T}) + \rho \parallel \mathbf{U} - \mathbf{A}^T\mathbf{Z} \parallel_F^2, \tag{19}$$

which can be further written as

$$\min_{\mathbf{u}^i} \mu(\mathbf{u}^{i^T} - \hat{\mathbf{P}}_i\mathbf{e}\log\mathbf{u}^{i^T}) + \rho \parallel \mathbf{u}^{i^T} - \mathbf{Z}^T\mathbf{a}_i \parallel_F^2 . \tag{20}$$

Using the similar approach as in (15-18), $\mathbf{u}_i^T$ is updated by

$$\mathbf{z}^{i^T} \leftarrow \text{prox}_{\tau(\cdot)}(\mathbf{u}^{i^T} - \delta(\mathbf{u}^{i^T} - \mathbf{Z}^T\mathbf{a}_i)), \tag{21}$$

where $\tau(\mathbf{u}^i) = \mu(\mathbf{u}^{i^T} - \hat{\mathbf{P}}_i\mathbf{e}\log\mathbf{u}^{i^T})$.

*Updating auxiliary variable **d**.* We can use the similar method as in (18) or (19) to update $\mathbf{d}$. Here we omit it.

Updating parameter set $\{\mathbf{P}_1, \cdots, \mathbf{P}_l, \hat{\mathbf{P}}_1, \cdots, \hat{\mathbf{P}}_m\}$. For fixed $\mathbf{Z}$ and $\mathbf{U}$, the update of each $\mathbf{P}_i$ or $\hat{\mathbf{P}}_i$ boils down to an OT problem, which can be solved via Sinkhorn iteration [Cuturi, 2013]. These steps are summarized in Algorithm 2, where we list the detailed iteration process for each $\mathbf{P}_i$. For each $\hat{\mathbf{P}}_i$, we can use the similar method to optimize it. The detailed iterative process is listed in Algorithm 3.

Convergence analysis. As pointed out by [Sandler and Lindenbaum, 2011], the alternate optimization process (Algorithm 1) generates a sequence of lower bounded non-increasing values for the objective of Problem (12), so the sequence of objectives converges. Meanwhile, according to [Janati *et al.*, 2018], we know that every accumulation point of the sequences of iterates of $\mathbf{Z}$, $\mathbf{U}$ and $\mathbf{P}_i$ (or $\hat{\mathbf{P}}_i$) is a generalized fixed point. For the convergence of Algorithm 2, it can be guaranteed by [Cuturi, 2013]. Here we omit it.

Algorithm 3 Sinkhorn iteration for optimizing $\hat{\mathbf{P}}_1, \cdots, \hat{\mathbf{P}}_m$

Input: data matrix $\mathbf{A}$ and coefficient matrix $\mathbf{Z}$
Output: $\hat{\mathbf{P}}_1, \cdots, \hat{\mathbf{P}}_m$
for $i = 1$ to m **do**
$\mathbf{K}_i = \exp(-\mathbf{C}_i/\gamma)$, where $\mathbf{C}_i$ is the ground metric matrix between $\mathbf{z}^{i^T}$ and $\varepsilon\mathbf{e}_n{}^T$, $\varepsilon > 0$.
repeat
$\mathbf{w}_i \leftarrow \mathbf{z}^{i^T}/\mathbf{K}\mathbf{v}_i$; $\mathbf{v}_i \leftarrow \varepsilon\mathbf{e}_n{}^T/\mathbf{K}^T\mathbf{w}_i$.
until convergence
$\hat{\mathbf{P}}_i \leftarrow (\hat{p}_{(i)jt})_{n\times n}$, where the (j,t)-th element of $\hat{\mathbf{P}}_i$ is $\hat{p}_{(i)jt} = w_{i(j)}k_{(i)jt}v_{i(t)}$ ($w_{i(j)}$ is the *j*-th element of $\mathbf{w}_i$).
end for

4 Theoretical Guarantee for Sinkhorn Regression

Tremendous studies have already been done on the statistical properties of Wasserstein Distances. For example, [Arras *et al.*, 2017] provided the bound to estimate the 2-Wasserstein distance between random variables which can be represented as linear combinations of independent random variables. [Frogner *et al.*, 2015] showed that Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space. However, theoretical Guarantee of the Sinkhorn distance loss minimization is still deficient. To bridge this gap, we will establish risk bounds of JWMR in this section. For the convenience, as in (8), let

$$\begin{aligned} L(\mathbf{Z}) &= \sum_{i=1}^{l} W_\gamma(h(\mathbf{Z}^T\mathbf{a}_i), h(\mathbf{y}^{i^T})), \\ \Omega(\mathbf{Z}) &= \sum_{i=1}^{m} W_\gamma(h(\mathbf{z}^{i^T}), \mathbf{d}), \end{aligned} \tag{22}$$

For a sequence $\mathcal{S} = \{(\mathbf{Z}_1, \mathbf{y}_1), \cdots, (\mathbf{Z}_N, \mathbf{y}_N)\}$ of i.i.d. training samples, we denote the empirical error and its expectation of $L(\mathbf{Z})$ as

$$R_e(\mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{Z}_i), R(\mathbf{Z}) = \mathcal{E}_\mathbf{Z}(L(\mathbf{Z})), \tag{23}$$

where $\mathcal{E}(\cdot)$ is the expectation operator.

To derive the bound of model (9), we need to introduce the following definition and Lemmas.

Definition 1. [Bartlett and Mendelson, 2002] Let $\mathcal{G}$ be a family of mapping from $\mathcal{Z}$ to $\mathcal{R}$, and $\mathcal{S} = (z_1, \cdots, z_N)$ a fixed sample from $\mathcal{Z}$. The empirical Rademacher complexity of $\mathcal{G}$ with regard to S is defined as

$$\mathfrak{R}_\mathcal{S}(\mathcal{G}) = \mathcal{E}_\sigma[\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^{n} \sigma_i g(z_i)] \tag{24}$$

where $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_N)$, with $\sigma_i's$ independent uniform random variables taking values in $\{+1, -1\}$. $\sigma_i's$ are called the Rademacher random variables. The Rademacher complexity is defines by taking expectation with respect to the samples S,

$$\mathfrak{R}_N(\mathcal{G}) = \mathcal{E}_{\mathcal{S}}[\hat{\mathfrak{R}}_S(\mathcal{G})]. \tag{25}$$

We define a space of loss function induced by the hypothesis space $\mathcal{Z}$ as

$$\mathcal{L} = \{\iota : (\mathbf{Z}, \mathbf{y}) :\to W_\gamma(h(\mathbf{Z}^T\mathbf{a}), h(\mathbf{y})), \mathbf{Z} \in \mathcal{Z}\}. \tag{26}$$

Lemma 2. [Frogner *et al.*, 2015] For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\iota \in \mathcal{L}$,

$$\mathcal{E}[\iota] - \hat{\mathcal{E}}_S[\iota] \leq 2\mathfrak{R}_N(\mathcal{L}) + \sqrt{\frac{M_C^2 \log(1/\delta)}{2N}} \tag{27}$$

with the constant $M_C = \max_{i,j} c_{i,j}$.

Assume that $\mathbf{Z}^*$ is the optimal solution of model (9), we have the following theorems:

Theorem 1. For any $\delta > 0$, with probability at least $1 - \delta$, it holds that

$$R_e(\mathbf{Z}^*) - R(\mathbf{Z}^*) \leq 16lKC_M\mathfrak{R}_N(\mathcal{H}^0) + 2lM_C\sqrt{\frac{\log(1/\delta)}{2N}} \tag{28}$$

where the constant $M_C = \max_{i,j} c_{i,j}$ and $\mathfrak{R}_N(\mathcal{H}^0)$ is the Rademacher complexity measure the complexity of the hypothesis space $\mathcal{H}^o$.

proof. Assume that the loss function $W_\gamma(h(\mathbf{Z}^T\mathbf{a}), h(\mathbf{y}))$ is preceded with a softmax layer

$$\mathcal{L} = \{\nu : (\mathbf{Z}, \mathbf{y}) :\to W_\gamma(s(h(\mathbf{Z}^{oT}\mathbf{a})), s(h(\mathbf{y}))), \mathbf{Z}^o \in \mathcal{Z}^o\} \tag{29}$$

We know that the function space can be expressed as:

$$\underbrace{\mathcal{H}^o \times \cdots \times \mathcal{H}^o}_{K \text{ copies}} \times \underbrace{\mathcal{I} \times \cdots \times \mathcal{I}}_{K \text{ copies}} \tag{30}$$

with $\mathcal{I}$ a singleton function space with only the identity map. Since $\hat{\mathfrak{R}}_S(\mathcal{I}) = 0$ (see (25)), it holds

$$\hat{\mathfrak{R}}_S(\mathcal{L}) \leq 8M_C(K\hat{\mathfrak{R}}_S(\mathcal{H}^o) + K\hat{\mathfrak{R}}_S(\mathcal{I})) = 8KM_C\hat{\mathfrak{R}}_S(\mathcal{H}^o). \tag{31}$$

Therefore, connecting Lemma 1, we have

$$\begin{aligned} & R_e(\mathbf{Z}^*) - R(\mathbf{Z}^*) \\ & \leq \sum_{i=1}^{l} \text{sup}_{\mathbf{Z}\in\mathcal{Z}} |r_e(\mathbf{Z}) - r(\mathbf{Z})| \\ & \leq \sum_{i=1}^{l} (16KC_M\mathfrak{R}_N(\mathcal{H}^0) + 2C_M\sqrt{\frac{\log(1/\delta)}{2N}}) \\ & < 16lKC_M\mathfrak{R}_N(\mathcal{H}^0) + 2lM_C\sqrt{\frac{\log(1/\delta)}{2N}}. \end{aligned} \tag{32}$$

This theorem shows that the Sinkhorn Regression estimator, as well as the empirical risk minimizer asymptotically reaches a $\sqrt{1/2N}$ speed of convergence under very weak hypotheses.

5 Experiments

In Section 3, we provided a general framework for Joint Sinkhorn matrix regression, which can be applied to many tasks such as feature selection, multi-task learning and image classification, *etc*. In this paper, we evaluate the performance of Sinkhorn regression using the feature selection task, which is important for many real-world applications, such as bioinformatics, text mining, *etc*. Here, $\mathbf{A} = [\mathbf{a}_i, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in \mathcal{R}^{m\times l}$ is a data matrix and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in R^{n\times c}$ is a label matrix. Our goal is to choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features.

5.1 Data Descriptions

We use five publicly available microarray data sets and one Mass Spectrometry (MS) datasets: ALLAML data set [Fodor, 1997], the malignant glioma (GLIOMA) data set [Nutt *et al.*, 2003], the human lung carcinomas (LUNG) data set [Bhattacharjee *et al.*, 2001], Human Carcinomas (Carcinomas) data set [Yang *et al.*, 2006], Prostate Cancer gene expression (Prostate-GE) data set [Singh *et al.*, 2002] for microarray data; and Prostate Cancer (Prostate-MS) [Petricoin III *et al.*, 2002] for MS data. To be fair, the Support Vector Machine (SVM) classifier is employed to these data sets, using 5-fold cross-validation for all compared methods.

5.2 Compared Methods

To validate the effectiveness of Sinkhorn regression for feature selection, we compare it with several classical feature selection methods. The compared algorithms are enumerated as follows.

1) F-statistic(F-s) [Ding and Peng, 2005]: it describes the statistically expected level of (usually) heterozygosity in a population; more specifically the expected degree of a reduction in heterozygosity when compared to Hardy–Weinberg expectation.

2) ReliefF(RF) [Kira and Rendell, 1992]: It relies entirely on statistical analysis and employs few heuristics and is less often foled.

3) mRMR [Sulaiman and Labadin, 2015]: A two-stage feature selection algorithm by combining mRMR and other more sophisticated feature selectors. 4) T-test: it determines if two sets of data are significantly different from each other

5) Information Gain (IG) [Raileanu and Stoffel, 2004]: A formal comparison of the behavior of two of the most popular split functions

6) Robust Feature Selection Based on $L_{2,1}$-Norms (RFS) [Nie *et al.*, 2010]: The $L_{2,1}$-norm is emphasizing on both loss function and regularization.

7) Clustering-Guided Sparse Structural Learning (CGSSL) [Li *et al.*, 2014]: An unsupervised feature selection framework by exploiting the cluster analysis and structural analysis with sparsity simultaneously.

8) Regularized Self-Representation (RSR) [Zhu *et al.*, 2015]: An unsupervised features election method by exploiting the self-representation ability of features.

Databases	RF	F-s	T-Test	IG	nRMR	RFS	CGSSL	RSR	Our method
ALLAML	90.36	89.11	92.86	93.21	93.21	95.89	94.27	93.89	**96.31**
GLIOMA	50	50	56	60	62	74	71	66	**74**
LUNG	91.68	87.7	89.22	93.1	92.61	93.63	92.77	89.36	**94.78**
Carcinom	79.88	65.48	49.9	85.09	78.22	91.38	88.32	87.29	**93.14**
Pro-GE	92.18	95.09	92.18	92.18	93.18	95.09	92.11	90.98	**95.16**
Pro-MS	76.41	98.89	95.56	98.89	95.42	**98.89**	97.32	97.31	98.71
Average	80.09	81.047	79.29	87.09	85.78	91.48	89.32	87.47	**92**

Table 1: Classification Accuracy (%) of SVM using 5-fold cross validation for all methods: average accuracy of top 20 features

Databases	RF	F-s	T-Test	IG	nRMR	RFS	CGSSL	RSR	Our method
ALLAML	95.89	96.07	94.29	95.71	94.46	97.32	95.21	94.27	**97.76**
GLIOMA	54	60	58	66	66	70	66	60	**70**
LUNG	93.63	91.63	90.66	95.1	94.12	96.07	93.50	92.14	**96.71**
Carcinom	90.24	83.33	68.91	89.65	87.92	93.66	91.21	91.31	**94.48**
Pro-GE	91.18	93.18	93.18	89.27	86.36	95.09	94.24	92.65	**95.89**
Pro-MS	89.93	98.89	94.44	98.89	93.14	100	97.39	93.87	**100**
Average	85.81	87.18	83.25	89.10	87	92.02	89.59	87.37	**92.47**

Table 2: Classification Accuracy (%) of SVM using 5-fold cross validation for all methods: average accuracy of top 80 features

λ	0.0001	0.001	0.01	0.05	0.1	1	5	10	100
ALLAML	88.75	90.92	92.12	92.87	97.76	**97.89**	96.65	92.27	86.81
GLIOMA	60	60	68	70	**70**	70	70	58	50
LUNG	92.17	94.32	96.11	**97.23**	96.71	95.32	95.11	83.98	72.23
Carcinom	93.80	96.47	**96.77**	95.21	94.48	93.73	93.23	85.98	81.21
Pro-GE	94.29	94.78	95.02	**96.73**	95.16	95.49	94.77	84.05	80.85
Pro-MS	97.96	98.87	99.12	99.45	**100**	100	98.74	88.70	79.92

Table 3: The influence of parameter λ to classification performance with top 20 features

5.3 Classification Accuracy Comparisons

In the experiments, all data sets are standardized to be zero-mean and normalized by standard deviation. SVM classifier has been individually performed on all data sets using 5-fold cross-validation. Our feature selection method is compared to several popularly used feature selection methods in bioinformatics, such as F-statistic, reliefF, mRMR, t-test, and information gain. Since the above data sets are for multiclass classification problem, L_1-SVM, HHSVM and other methods that were designed for binary classification are not compared Table 1 and 2 show the detailed experimental results using SVM classifier. The average accuracy for each feature selection approach is calculated using the top 20 and top 80 features. It can be seen that our approaches obviously outperform most of methods significantly. Specifically, with top 20 features, our method is around 0%-2% better than other methods all six data sets. With top 80 features, our method achieve an improvement of 0.45% than the second best method (RFS) on average accuracy.

Like many other feature selection algorithms, our proposed Sinkhorn regression also requires several parameters λ, γ and μ to be set in advance. For the results reported in the above subsection, we do not tune the parameter γ and μ and only set them as: $\gamma = 0.01$ and $\mu = 0.1$. Better results may be achieved with tuning it. In this subsection, we will discuss sensitivity of parameter λ. Here, we take the top 20 features as an example on all data sets. The detailed results are shown in Table 3. It can be found the best results of our method mainly lie in the interval $[0.01 \;\; 0.5]$. But in Table 1 and 2, we choose $\lambda = 0.1$.

6 Conclusions

In this paper, we proposed a new robust matrix regression method with emphasizing joint Sinkhorn distances minimization on both loss function and regularization. The Sinkhorn distances based loss function is robust to noise in data points. Meanwhile, Sinkhorn based regularizer can promote parameters that are close. In addition, the proposed Sinkhorn regression is extended to the unbalanced formulation which does not rely on distribution features of data. We provided an efficient algorithm to solve the proposed new model and described its generalization bound from the statistical viewpoint. Our method has been applied into feature selection task. Extensive empirical studies on six standard data sets demonstrated that the proposed algorithm works much more robustly than some existing methods.

References

[Arras *et al.*, 2017] Benjamin Arras, Ehsan Azmoodeh, Guillaume Poly, and Yvik Swan. A bound on the 2-wasserstein distance between linear combinations of independent random variables. *arXiv preprint arXiv:1704.01376*, 2017.

[Bartlett and Mendelson, 2002] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002.

[Bhattacharjee *et al.*, 2001] Arindam Bhattacharjee, William G Richards, Jane Staunton, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98(24):13790–13795, 2001.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

[Ding and Peng, 2005] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *JBCB*, 3(02):185–205, 2005.

[Flamary *et al.*, 2016] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, pages 1–23, 2016.

[Fodor, 1997] Stephen PA Fodor. Massively parallel genomics. *Science*, 277(5324):393, 1997.

[Frogner *et al.*, 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *NeurIPS*, pages 2053–2061, 2015.

[He *et al.*, 2011] Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE TPAMI*, 33(8):1561–1576, 2011.

[Janati *et al.*, 2018] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. *arXiv preprint arXiv:1805.07833*, 2018.

[Jia *et al.*, 2012] Kui Jia, Tsung-Han Chan, and Yi Ma. Robust and practical face recognition via structured sparsity. In *ECCV*, pages 331–344. Springer, 2012.

[Kira and Rendell, 1992] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier, 1992.

[Kolouri and Rohde, 2015] Soheil Kolouri and Gustavo K Rohde. Transport-based single frame super resolution of very low resolution face images. In *CVPR*, pages 4876–4884, 2015.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966, 2015.

[Li *et al.*, 2013] Xiao-Xin Li, Dao-Qing Dai, Xiao-Fei Zhang, and Chuan-Xian Ren. Structured sparse error coding for face recognition with occlusion. *IEEE TIP*, 22(5):1889–1900, 2013.

[Li *et al.*, 2014] Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou, et al. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE TKDE*, 26(9):2138–2150, 2014.

[Liu *et al.*, 2010] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[Lu *et al.*, 2012] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, et al. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, pages 347–360. Springer, 2012.

[Naseem *et al.*, 2010] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Linear regression for face recognition. *IEEE TPAMI*, 32(11):2106–2112, 2010.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. In *NeurIPS*, pages 1813–1821, 2010.

[Nutt *et al.*, 2003] Catherine L Nutt, DR Mani, Rebecca A Betensky, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research*, 63(7):1602–1607, 2003.

[Ozolek *et al.*, 2014] John A Ozolek, Akif Burak Tosun, Wei Wang, Cheng Chen, et al. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical image analysis*, 18(5):772–780, 2014.

[Petricoin III *et al.*, 2002] Emanuel F Petricoin III, David K Ornstein, Cloud P Paweletz, Ali Ardekani, Paul S Hackett, Ben A Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, Kamillah Wood, et al. Serum proteomic patterns for detection of prostate cancer. *JNCI*, 94(20):1576–1578, 2002.

[Raileanu and Stoffel, 2004] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *AMAI*, 41(1):77–93, 2004.

[Rolet *et al.*, 2016] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638, 2016.

[Sandler and Lindenbaum, 2011] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE TPAMI*, 33(8):1590–1602, 2011.

[Singh *et al.*, 2002] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[Sommerfeld *et al.*, 2018] Max Sommerfeld, Jörn Schrieber, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *arXiv preprint arXiv:1802.05570*, 2018.

[Sulaiman and Labadin, 2015] Muhammad Aliyu Sulaiman and Jane Labadin. Feature selection based on mutual information. In *CITA*, pages 1–6. IEEE, 2015.

[Yang *et al.*, 2006] Kun Yang, Zhipeng Cai, Jianzhong Li, and Guohui Lin. A stable gene selection in microarray data analysis. *BMC bioinformatics*, 7(1):228, 2006.

[Yang *et al.*, 2011] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *CVPR 2011*, pages 625–632. IEEE, 2011.

[Yang *et al.*, 2012] Jian Yang, Lei Zhang, Yong Xu, and Jing-yu Yang. Beyond sparsity: The role of l1-optimizer in pattern classification. *Pattern Recognition*, 45(3):1104–1118, 2012.

[Yang *et al.*, 2013] Jian Yang, Delin Chu, Lei Zhang, Yong Xu, and Jingyu Yang. Sparse representation classifier steered discriminative projection with applications to face recognition. *IEEE TNNLS*, 24(7):1023–1035, 2013.

[Yang *et al.*, 2017] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE TPAMI*, 39(1):156–171, 2017.

[Zhu *et al.*, 2015] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, et al. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.