2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

A Distributionally Robust Optimization Approach for Multivariate Linear Regression under the Wasserstein Metric *

Ruidi Chen[1] and Ioannis Ch. Paschalidis[2]

***Abstract*— We present a *Distributionally Robust Optimization (DRO)* approach for *Multivariate Linear Regression (MLR)*, where multiple correlated response variables are to be regressed against a common set of predictors. We develop a regularized MLR formulation that is robust to large perturbations in the data, where the regularizer is the dual norm of the regression coefficient matrix in the sense of a newly defined matrix norm. We establish bounds on the prediction bias of the solution, offering insights on the role of the regularizer in controlling the prediction error. Experimental results show that, compared to a number of popular MLR methods, our approach leads to a lower out-of-sample *Mean Squared Error (MSE)* in various scenarios.**

I. INTRODUCTION

We are interested in the problem of regressing multiple correlated responses against a common set of predictors, which we call *Multivariate Linear Regression (MLR)*. This term is distinct from *multiple linear regression*, where only a scalar response with more than one predictors is involved. MLR has found applications in econometrics [1], health care [2], and finance [3]. It is useful when multiple measurements for a single individual are available [4], or the valuation of a group of interdependent variables is of interest [5]. It can also be used for multiple-task learning [6], where a set of related tasks are to be learned simultaneously.

We assume the following model for the MLR problem:

$$\mathbf{y} = \mathbf{B}'\mathbf{x} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_K)$ is the vector of K responses, potentially correlated with each other; $\mathbf{x} = (x_1, \ldots, x_p)$ is the vector of p predictors; $\mathbf{B} = (B_{ij})_{i=1,\ldots,p}^{j=1,\ldots,K}$ is the $p \times K$ matrix of coefficients, the j-th column of which describes the dependency of y_j on the predictors; $\boldsymbol{\epsilon}$ is the random error and prime denotes transpose. Suppose we observe N realizations of the data, denoted by $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), \mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$. *Ordinary Least Squares (OLS)* solves the regression coefficients by minimizing the sum of squared errors, which is equivalent to regressing each response variable against the predictors independently. It does not take into account the potential correlation existing among the responses, and is vulnerable to large perturbations in the data. Its performance is significantly degraded when the predictors are highly correlated or p is relatively large [7].

A class of methods that are used to overcome the aforementioned problems is called linear factor regression, where the response $\mathbf{y}$ is regressed against a small number of linearly transformed predictors (factors). Examples include reduced rank regression [8], principal components regression [9], and *Factor Estimation and Selection (FES)* [7]. Another type of methods apply multivariate shrinkage by either estimating a linear transformation of the OLS predictions [5], or solving a regularized MLR problem, e.g., ridge regression [10], and FES [7] whose regularizer is the coefficient matrix's Ky Fan norm defined as the sum of its singular values.

None of the aforementioned methods, however, explicitly take into account the robustness of the model, and could result in estimates that are vulnerable to adversarial perturbations in the data. The multivariate extension of ridge regression, which penalizes the trace of $\mathbf{B}'\mathbf{B}$, hedges against large noise to some extent, but is criticized for not offering interpretable models because of the dense coefficient estimates [7].

In this paper, we address this problem by adopting a *Distributionally Robust Optimization (DRO)* formulation that minimizes the worst-case expected loss within a probabilistic ambiguity set defined by the Wasserstein metric [11, 12]. This approach induces robustness by hedging against a set of probability distributions, and has been extensively studied in the single-response scenario [13, 14, 15, 16]. However, there is no work examining the multivariate DRO problem, which is a nontrivial extension of the prior work in light of the correlation between responses and the geometrical structure of the coefficient matrix. We will fill this gap and establish the connection between robustness and regularization in the multivariate scenario by defining a new notion of norm on the regression coefficient matrix.

To the best of our knowledge, we are the first to study the distributionally robust MLR problem, without imposing any assumption on the correlation structure of the response variables. We relax the DRO-MLR formulation into a convex regularized regression problem, with the regularizer being the dual norm of the coefficient matrix, in the sense of a newly defined matrix norm that scalarizes each column by the sum of the absolute values of its elements. This model is completely optimization-based, and avoids the need of explicitly modeling the relationship between responses. It is computationally more efficient to solve, and is more robust

* Research partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under MURI grant N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, by the Boston University Digital Health Initiative, and by the Boston University Center for Information and Systems Engineering.

[1]Ruidi Chen is with Division of Systems Engineering, Boston University, Boston, MA 02446, USA. rchen15@bu.edu

[2]Ioannis Ch. Paschalidis is with Dept. of Electrical and Computer Engineering, Division of Systems Engineering, and Dept. of Biomedical Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, USA. yannisp@bu.edu, http://sites.bu.edu/paschalidis/.

to outliers than other MLR models.

The rest of the paper is organized as follows. In Section II, we develop the DRO-MLR formulation and introduce the new matrix norm that is involved in its relaxation. Section III establishes the out-of-sample performance guarantee. The numerical experimental results are presented in Section IV. We conclude the paper in Section V.

Notational conventions: We use boldfaced lowercase letters to denote vectors, ordinary lowercase letters to denote scalars, boldfaced uppercase letters to denote matrices, and calligraphic capital letters to denote sets. $\mathbb{E}$ denotes expectation and $\mathbb{P}$ probability of an event. All vectors are column vectors. For space saving reasons, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. We use prime to denote the transpose, $\|\cdot\|_p$ for the ℓ_p norm with $p \geq 1$, and $\|\cdot\|$ for the general vector norm that satisfies the following properties: (i) $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = \mathbf{0}$; (ii) $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, for any scalar a; (iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$; (iv) $\|\mathbf{x}\| = \||\mathbf{x}|\|$, where $|\mathbf{x}| = (|x_1|, \ldots, |x_{\dim(\mathbf{x})}|)$; and (v) $\|(\mathbf{x}, \mathbf{0})\| = \|\mathbf{x}\|$, for an arbitrarily long vector $\mathbf{0}$. Note that any $\mathbf{W}$-weighted ℓ_p norm defined as $\|\mathbf{x}\|_p^{\mathbf{W}} \triangleq \left((|\mathbf{x}|^{p/2})'\mathbf{W}|\mathbf{x}|^{p/2}\right)^{1/p}$ with a positive definite $\mathbf{W}$ satisfies the above conditions, where $|\mathbf{x}|^{p/2} = (|x_1|^{p/2}, \ldots, |x_{\dim(\mathbf{x})}|^{p/2})$. Finally, $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$ defined as $\|\boldsymbol{\theta}\|_* \triangleq \sup_{\|\mathbf{z}\| \leq 1} \boldsymbol{\theta}'\mathbf{z}$, and $\mathbf{I}_K$ denotes the K-dimensional identity matrix.

II. Formulation

In this section we introduce the Wasserstein DRO formulation for MLR, and present a matrix norm interpretation that resembles the single-response scenario developed in [13].

A. The Wasserstein DRO Formulation

The Wasserstein DRO formulation for MLR minimizes the following worst-case expected loss:

$$\inf_{\mathbf{B}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}} \|\mathbf{y} - \mathbf{B}'\mathbf{x}\|_1, \tag{1}$$

where $\mathbb{Q}$ is the probability distribution of the data $(\mathbf{x}, \mathbf{y})$, belonging to a set Ω defined as

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\},$$

where $\mathcal{Z}$ is the set of possible values for $(\mathbf{x}, \mathbf{y})$; $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$; ϵ is a pre-specified positive constant that measures the size of the set Ω; $\hat{\mathbb{P}}_N$ is the empirical distribution that assigns equal probability to each observed sample; $W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the order-1 Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ defined as

$$W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \triangleq \inf_{\Pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} \Big\{ \int_{\mathcal{Z} \times \mathcal{Z}} \|\mathbf{z}_1 - \mathbf{z}_2\|\ \Pi\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \Big\}, \tag{2}$$

where $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$, $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2)$; and Π is the joint distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$ with marginals being $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, respectively; $\|\cdot\|$ could be any vector norm which measures the cost of transporting the probability mass.

Define the loss function $h_{\tilde{\mathbf{B}}}(\mathbf{z}) \triangleq \|\tilde{\mathbf{B}}\mathbf{z}\|_1$, where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, and $\tilde{\mathbf{B}} = (-\mathbf{B}', \mathbf{I}_K)$. To make the inner supremum of (1) finite, we first observe that, for any $\mathbb{Q} \in \Omega$,

$$\begin{aligned}
&\Big|\mathbb{E}^{\mathbb{Q}}\big[h_{\tilde{\mathbf{B}}}(\mathbf{z})\big] - \mathbb{E}^{\hat{\mathbb{P}}_N}\big[h_{\tilde{\mathbf{B}}}(\mathbf{z})\big]\Big| \\
=&\Big|\int_{\mathcal{Z}} h_{\tilde{\mathbf{B}}}(\mathbf{z}_1)\mathbb{Q}(d\mathbf{z}_1) - \int_{\mathcal{Z}} h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\hat{\mathbb{P}}_N(d\mathbf{z}_2)\Big| \\
=&\Big|\int_{\mathcal{Z}} h_{\tilde{\mathbf{B}}}(\mathbf{z}_1)\int_{\mathcal{Z}} \Pi_0(d\mathbf{z}_1, d\mathbf{z}_2) - \int_{\mathcal{Z}} h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\int_{\mathcal{Z}} \Pi_0(d\mathbf{z}_1, d\mathbf{z}_2)\Big| \\
\leq&\int_{\mathcal{Z} \times \mathcal{Z}} \big|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\big|\Pi_0(d\mathbf{z}_1, d\mathbf{z}_2),
\end{aligned} \tag{3}$$

where Π_0 is the joint distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$ that achieves the optimal value of (2). Comparing (3) with (2), we see that to establish a connection between the expected loss difference and the Wasserstein distance, bounding the following *growth rate* of the loss is the key. Define the *Growth Rate (GR)* of $h_{\tilde{\mathbf{B}}}(\cdot)$ as:

$$\begin{aligned}
\mathrm{GR}\big(h_{\tilde{\mathbf{B}}}\big) &\triangleq \frac{\big|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\big|}{\|\mathbf{z}_1 - \mathbf{z}_2\|} \\
&= \frac{\big|\|\tilde{\mathbf{B}}\mathbf{z}_1\|_1 - \|\tilde{\mathbf{B}}\mathbf{z}_2\|_1\big|}{\|\mathbf{z}_1 - \mathbf{z}_2\|} \\
&\leq \frac{\|\tilde{\mathbf{B}}(\mathbf{z}_1 - \mathbf{z}_2)\|_1}{\|\mathbf{z}_1 - \mathbf{z}_2\|}, \quad \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z},
\end{aligned}$$

where in the last step we use the reverse triangle inequality. We would like to derive an upper bound for $\mathrm{GR}(h_{\tilde{\mathbf{B}}})$ that is independent of the data $\mathbf{z}_1$ and $\mathbf{z}_2$. To this end, it is desired to bound $\|\tilde{\mathbf{B}}(\mathbf{z}_1 - \mathbf{z}_2)\|_1$ in terms of $\|\mathbf{z}_1 - \mathbf{z}_2\|$. The following two corollaries serve this purpose.

Corollary II.1. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, we have*

$$\|\mathbf{A}\mathbf{x}\|_1 \leq \|\mathbf{x}\| \sum_{i=1}^m \|\mathbf{a}_i\|_*,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$.

Proof. By the Cauchy-Schwarz inequality, we have:

$$\|\mathbf{A}\mathbf{x}\|_1 = \sum_{i=1}^m |\mathbf{a}_i'\mathbf{x}| \leq \sum_{i=1}^m \|\mathbf{x}\|\|\mathbf{a}_i\|_*.$$

□

Corollary II.2. *Given an $m \times n$ matrix $\mathbf{A} = (a_{ij})_{i=1,\ldots,m}^{j=1,\ldots,n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the following holds:*

$$\|\mathbf{A}\mathbf{x}\|_1 \leq \|\mathbf{v}\|_*\|\mathbf{x}\|,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and $\mathbf{v} = (v_1, \ldots, v_n)$, with $v_j = \sum_{i=1}^m |a_{ij}|$.

Proof.

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_1 &= \Big|\sum_{j=1}^n a_{1j}x_j\Big| + \ldots + \Big|\sum_{j=1}^n a_{mj}x_j\Big| \\
&\leq |x_1|\sum_{i=1}^m |a_{i1}| + \ldots + |x_n|\sum_{i=1}^m |a_{in}| \\
&\triangleq \mathbf{v}'\bar{\mathbf{x}},
\end{aligned}$$

where $\mathbf{v} = (v_1, \ldots, v_n)$, with $v_j = \sum_{i=1}^m |a_{ij}|$, and $\bar{\mathbf{x}} = (|x_1|, \ldots, |x_n|)$. Thus,

$$\|\mathbf{A}\mathbf{x}\|_1 \leq \|\mathbf{v}\|_* \|\bar{\mathbf{x}}\| = \|\mathbf{v}\|_* \|\mathbf{x}\|.$$

□

Remark 2.1: Corollaries II.1 and II.2 provide two forms of bounds for the ℓ_1 norm of the matrix-vector product. Notice that $\mathbf{v} = \sum_{i=1}^m |\mathbf{a}_i|$, where the $|\cdot|$ is applied element-wise to $\mathbf{a}_i$, and therefore,

$$\|\mathbf{v}\|_* = \Big\| \sum_{i=1}^m |\mathbf{a}_i| \Big\|_* \leq \sum_{i=1}^m \|\mathbf{a}_i\|_*,$$

implying that Corollary II.2 gives a tighter bound.

We now proceed to obtain a tractable upper bound to the inner supremum of (1). Using Corollary II.1, (3) can be further bounded by:

$$\begin{aligned}
&\Big|\mathbb{E}^{\mathbb{Q}}\big[h_{\tilde{\mathbf{B}}}(\mathbf{z})\big] - \mathbb{E}^{\hat{\mathbb{P}}_N}\big[h_{\tilde{\mathbf{B}}}(\mathbf{z})\big]\Big| \\
&\leq \int_{\mathcal{Z}\times\mathcal{Z}} \big|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\big| \Pi_0(d\mathbf{z}_1, d\mathbf{z}_2) \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \frac{\big|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)\big|}{\|\mathbf{z}_1 - \mathbf{z}_2\|} \|\mathbf{z}_1 - \mathbf{z}_2\| \Pi_0(d\mathbf{z}_1, d\mathbf{z}_2) \\
&\leq \Big(\sum_{i=1}^K \|\mathbf{b}_i\|_*\Big) \int_{\mathcal{Z}\times\mathcal{Z}} \|\mathbf{z}_1 - \mathbf{z}_2\| \Pi_0(d\mathbf{z}_1, d\mathbf{z}_2) \\
&= \Big(\sum_{i=1}^K \|\mathbf{b}_i\|_*\Big) W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \\
&\leq \epsilon \sum_{i=1}^K \|\mathbf{b}_i\|_*,\ \forall \mathbb{Q} \in \Omega,
\end{aligned}$$

where $\mathbf{b}_i = (-B_{1i}, \ldots, -B_{pi}, \mathbf{e}_i)$ is the i-th row of $\tilde{\mathbf{B}}$, with $\mathbf{e}_i$ the i-th unit vector in $\mathbb{R}^K$. The above derivation implies that the inner supremum of (1) can be upper bounded by:

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}} \|\mathbf{y} - \mathbf{B}'\mathbf{x}\|_1 \leq \mathbb{E}^{\hat{\mathbb{P}}_N}\big[h_{\tilde{\mathbf{B}}}(\mathbf{z})\big] + \epsilon \sum_{i=1}^K \|\mathbf{b}_i\|_*,$$

which directly yields the following relaxation to (1):

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i\|_1 + \epsilon \sum_{i=1}^K \|\mathbf{b}_i\|_*. \tag{4}$$

Problem (4) is equivalent to solving a single-response regularized regression formulation, which is a relaxation to the single-response Wasserstein DRO problem [13], for each of the K responses separately. Though the regularizer guarantees robustness to large perturbations, the correlation between responses is still not explored. Using a similar derivation, Corollary II.2 yields the following relaxation to (1):

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i\|_1 + \epsilon \|\mathbf{v}\|_*, \tag{5}$$

where $\mathbf{v} \triangleq (v_1, \ldots, v_p, 1, \ldots, 1)$, with $v_i = \sum_{j=1}^K |B_{ij}|$, i.e., v_i is a condensed representation of the coefficients for predictor i through summing over the K coordinates. Formulation (5) cannot be decomposed into K subproblems due to the entangling of coefficients in the regularization term. It is though computationally efficient to solve due to the convexity of the loss and the regularizer. According to Remark 2.1, it serves as a tighter relaxation than (4).

The growth rate analysis gives us an intuitive derivation of the relaxations. We next present a more rigorous proof for formulation (5), based on a convex program reformulation developed in [12]. When the loss function is convex in the data that resides in a closed and convex set, Theorem 6.3 in [12] shows that the worst-case expected loss can be upper bounded by

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[\|\mathbf{y} - \mathbf{B}'\mathbf{x}\|_1] \leq \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i\|_1 + \kappa\epsilon, \tag{6}$$

where $\kappa = \sup\{\|\boldsymbol{\theta}\|_* : h^*_{\tilde{\mathbf{B}}}(\boldsymbol{\theta}) < \infty\}$, and $h^*_{\tilde{\mathbf{B}}}(\cdot)$ is the conjugate function of $h_{\tilde{\mathbf{B}}}(\cdot)$ defined as $h^*_{\tilde{\mathbf{B}}}(\boldsymbol{\theta}) \triangleq \sup_{\mathbf{z}} \{\boldsymbol{\theta}'\mathbf{z} - h_{\tilde{\mathbf{B}}}(\mathbf{z})\}$. In the next theorem we will build a connection between κ and the regression coefficients $\tilde{\mathbf{B}}$, and try to recover (5) from the upper bound given in (6).

Theorem II.3. *Define $\kappa = \sup\{\|\boldsymbol{\theta}\|_* : h^*_{\tilde{\mathbf{B}}}(\boldsymbol{\theta}) < \infty\}$, where $\|\cdot\|_*$ is the dual norm of the norm that is used to define the Wasserstein metric in (2), and $h^*_{\tilde{\mathbf{B}}}(\cdot)$ is the conjugate function of $h_{\tilde{\mathbf{B}}}(\cdot)$. When the loss function is $h_{\tilde{\mathbf{B}}}(\mathbf{z}) = \|\tilde{\mathbf{B}}\mathbf{z}\|_1$, we have $\kappa = \|\mathbf{v}\|_*$, where $\mathbf{v} \in \mathbb{R}^{p+K}$, and the j-th element of $\mathbf{v}$ is the sum of absolute values of the j-th column of $\tilde{\mathbf{B}}$.*

Proof. Consider the following optimization problem:

$$\max_{\mathbf{z}} \quad \boldsymbol{\theta}'\mathbf{z} - \|\tilde{\mathbf{B}}\mathbf{z}\|_1.$$

We can translate it into the linear programming problem:

$$\begin{aligned}
\max_{\mathbf{z}, r_i} \quad & \boldsymbol{\theta}'\mathbf{z} - r_1 - \ldots - r_K \\
\text{s.t.} \quad & r_1 - \mathbf{b}_1'\mathbf{z} \geq 0, \\
& r_1 + \mathbf{b}_1'\mathbf{z} \geq 0, \\
& \quad \vdots \\
& r_K - \mathbf{b}_K'\mathbf{z} \geq 0, \\
& r_K + \mathbf{b}_K'\mathbf{z} \geq 0,
\end{aligned}$$

where $\mathbf{b}_i$ is the i-th row of $\tilde{\mathbf{B}}$. Form its dual using dual variables $q_i, s_i, i = 1, \ldots, K$:

$$\begin{aligned}
\min_{q_i, s_i} \quad & 0 \\
\text{s.t.} \quad & \mathbf{b}_1(s_1 - q_1) + \ldots + \mathbf{b}_K(s_K - q_K) = \boldsymbol{\theta}, \\
& q_i + s_i = -1, \forall i = 1, \ldots, K, \\
& q_i, s_i \leq 0.
\end{aligned}$$

In order to make the optimal value of the primal problem finite, as required by the definition of κ, the dual needs to be feasible. From the first constraint of the dual, we have the following:

$$\|\boldsymbol{\theta}\|_* = \|\tilde{\mathbf{B}}'\mathbf{w}\|_* = \|(\mathbf{t}_1'\mathbf{w}, \ldots, \mathbf{t}_{p+K}'\mathbf{w})\|_*,$$

where $\mathbf{w} = (s_1 - q_1, \ldots, s_K - q_K)$, $\mathbf{t}_i$ is the i-th column of $\tilde{\mathbf{B}}$, for $i = 1, \ldots, p + K$.

Write $\mathbf{v} \triangleq (v_1, \ldots, v_p, 1, \ldots, 1)$, with $v_i = \sum_{j=1}^{K} |B_{ij}|$. The last two constraints of the dual imply that $|s_i - q_i| \leq 1, \forall i$, which yields the following:

$$|\mathbf{t}_i'\mathbf{w}| = \Big|\sum_{j=1}^{K} -B_{ij}(s_j - q_j)\Big| \leq \sum_{j=1}^{K} |B_{ij}| = v_i, \forall i = 1, \ldots, p,$$

and,

$$|\mathbf{t}_{p+i}'\mathbf{w}| = |s_i - q_i| \leq 1, \forall i = 1, \ldots, K.$$

Therefore,

$$\|\boldsymbol{\theta}\|_* \leq \|(v_1, \ldots, v_p, 1, \ldots, 1)\|_* = \|\mathbf{v}\|_*,$$

which leads to the conclusion that $\kappa = \sup\{\|\boldsymbol{\theta}\|_* : h^*_{\tilde{\mathbf{B}}}(\boldsymbol{\theta}) < \infty\} = \|\mathbf{v}\|_*$. □

Plugging the value of κ into (6), we arrive at the relaxation (5). Note that the regularizer depends on the norm that is used to define the Wasserstein metric, and the regularization coefficient coincides with the size of the underlying distributional ambiguity set. Such a regularized regression formulation stems from the basic DRO problem, ensuring its robustness to larger perturbations, and providing a fundamental way of preventing overfitting in the multiple-response setting.

B. A New Perspective on the Formulation

In this subsection we will present a matrix norm interpretation for formulation (5). Different from the commonly used matrix norm definitions in the literature, e.g., the vector norm-induced matrix norm $\|\mathbf{A}\| \triangleq \max_{\|\mathbf{x}\| \leq 1} \|\mathbf{A}\mathbf{x}\|$, the entrywise norm that treats the matrix as a vector, and the Schatten norm that defines the norm on the vector of singular values [17], we adopt a new notion of matrix norm that summarizes each column by its absolute sum. We will call it the *Column Matrix Norm*.

Definition 1 (Column Matrix Norm)**.** *For any $m \times n$ matrix $\mathbf{A} = (a_{ij})_{i=1,\ldots,m}^{j=1,\ldots,n}$, define its column matrix norm as:*

$$\|\mathbf{A}\| \triangleq \|\mathbf{v}\|,$$

where $\|\cdot\|$ could be any vector norm operator, and $\mathbf{v} = (v_1, \ldots, v_n)$, with $v_j = \sum_{i=1}^{m} |a_{ij}|$. We write $\|\mathbf{A}\|_p$ to denote the ℓ_p-norm induced column matrix norm.

We note that the column matrix norm depends on the structure of the matrix, and transposing a matrix changes its norm. For example, given $\mathbf{A} \in \mathbb{R}^{n \times 1}$, $\|\mathbf{A}\|_p = \|\mathbf{a}\|_1$, $\|\mathbf{A}'\|_p = \|\mathbf{a}\|_p$, where $\mathbf{a}$ represents the vectorization of $\mathbf{A}$. We next show that the column matrix norm is a valid norm. It is easy to verify that:

1) $\|\mathbf{A}\| \geq 0$.
2) $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$.
3) $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$.
4) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

The column matrix norm also satisfies the following *sub-multiplicative* property:

$$\|\mathbf{A}\mathbf{B}\|_p \leq \|\mathbf{A}\|_q \|\mathbf{B}\|_p,$$

for $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times r}$, and $p, q \geq 1$.

Next we will reformulate (5) using the column matrix norm. Notice that the regularizer of (5) is just the dual norm of $\tilde{\mathbf{B}}$. Thus, it is equivalent to:

$$\inf_{\tilde{\mathbf{B}}} \frac{1}{N} \sum_{i=1}^{N} \|\tilde{\mathbf{B}}\mathbf{z}_i\|_1 + \epsilon\|\tilde{\mathbf{B}}\|_*, \tag{7}$$

which is in the same form as formulation (10) in [13], where the Wasserstein DRO relaxation for the single-response case was presented. This reformulation allows us to explore the predictive performance of the solution to (7), which enables a quantitative characterization of the performance of the two relaxations (4) and (5), and will be discussed in Section III.

III. Predictive Performance

In this section we study the out-of-sample predictive performance of the solutions to (4) and (5) using Rademacher complexity [18], which is a measurement of the complexity of a class of functions. Though the derivation technique is not new, see [13], the resulting bounds are informative for understanding the role of the regularizer, enabling a comparison between (4) and (5) in terms of their prediction biases. We first make the following assumptions that are essential for deriving the bounds.

Assumption A. *The norm of the data $(\mathbf{x}, \mathbf{y})$ is bounded above almost surely, i.e., $\|(\mathbf{x}, \mathbf{y})\| \leq R$.*

Assumption B. $\sum_{i=1}^{K} \|\mathbf{b}_i\|_* \leq \bar{B}_1$*, where $\mathbf{b}_i$ is the i-th row of $\tilde{\mathbf{B}}$.*

Assumption C. $\|\tilde{\mathbf{B}}\|_* \leq \bar{B}_2$.

Under Assumptions A and B, Corollary II.1 yields that

$$\|\tilde{\mathbf{B}}\mathbf{z}\|_1 \leq \|\mathbf{z}\| \sum_{i=1}^{K} \|\mathbf{b}_i\|_* \leq R\bar{B}_1.$$

Similarly, under Assumptions A and C, Corollary II.2 yields the following:

$$\|\tilde{\mathbf{B}}\mathbf{z}\|_1 \leq \|\mathbf{z}\|\|\tilde{\mathbf{B}}\|_* \leq R\bar{B}_2.$$

With the above results, the idea is to bound the out-of-sample prediction error using the empirical *Rademacher complexity* $\mathcal{R}_N(\cdot)$ of the following class of loss functions:

$$\mathcal{H} = \{\mathbf{z} \mapsto h_{\tilde{\mathbf{B}}}(\mathbf{z}) : h_{\tilde{\mathbf{B}}}(\mathbf{z}) = \|\tilde{\mathbf{B}}\mathbf{z}\|_1\},$$

which is defined as:

$$\mathcal{R}_N(\mathcal{H}) \triangleq \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{2}{N} \left|\sum_{i=1}^{N} \sigma_i h_{\tilde{\mathbf{B}}}(\mathbf{z}_i)\right| \middle| \mathbf{z}_1, \ldots, \mathbf{z}_N\right],$$

where $\sigma_1, \ldots, \sigma_N$ are i.i.d. uniform random variables on $\{1, -1\}$.

Lemma III.1. *Under Assumptions A and B,*

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2\bar{B}_1 R}{\sqrt{N}}.$$

Under Assumptions A and C,

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2\bar{B}_2 R}{\sqrt{N}}.$$

Lemma III.1 can be proved by plugging the corresponding upper bounds on the loss functions into Lemma 3.2 of [13]. Using the Rademacher complexity of the loss functions, the out-of-sample prediction biases of the solutions to (4) and (5) can be bounded by applying Theorem 8 in [18].

Theorem III.2. *Suppose the solution to (4) is $\hat{\mathbf{B}}$. Under Assumptions A and B, for any $0 < \delta < 1$, with probability at least $1-\delta$ with respect to the sampling,*

$$\mathbb{E}[\|\mathbf{y}-\hat{\mathbf{B}}'\mathbf{x}\|_1] \leq \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\frac{2\bar{B}_1 R}{\sqrt{N}}+\bar{B}_1 R\sqrt{\frac{8\log(\frac{2}{\delta})}{N}},$$

and for any $\zeta > \frac{2\bar{B}_1 R}{\sqrt{N}} + \bar{B}_1 R\sqrt{\frac{8\log(2/\delta)}{N}}$,

$$\mathbb{P}\Big(\|\mathbf{y}-\hat{\mathbf{B}}'\mathbf{x}\|_1 \geq \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\zeta\Big) \leq \frac{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\frac{2\bar{B}_1 R}{\sqrt{N}}+\bar{B}_1 R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\zeta}.$$

Theorem III.3. *Suppose the solution to (5) is $\hat{\mathbf{B}}$. Under Assumptions A and C, for any $0 < \delta < 1$, with probability at least $1-\delta$ with respect to the sampling,*

$$\mathbb{E}[\|\mathbf{y}-\hat{\mathbf{B}}'\mathbf{x}\|_1] \leq \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\frac{2\bar{B}_2 R}{\sqrt{N}}+\bar{B}_2 R\sqrt{\frac{8\log(\frac{2}{\delta})}{N}},$$

and for any $\zeta > \frac{2\bar{B}_2 R}{\sqrt{N}} + \bar{B}_2 R\sqrt{\frac{8\log(2/\delta)}{N}}$,

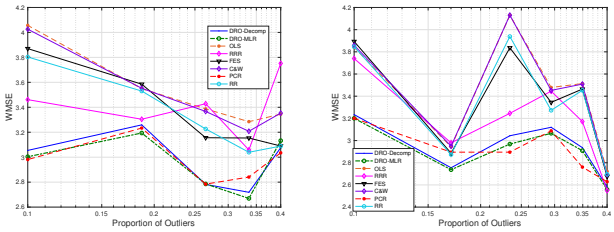
$$\mathbb{P}\Big(\|\mathbf{y}-\hat{\mathbf{B}}'\mathbf{x}\|_1 \geq \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\zeta\Big) \leq \frac{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\frac{2\bar{B}_2 R}{\sqrt{N}}+\bar{B}_2 R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{B}}'\mathbf{x}_i\|_1+\zeta}.$$

Remark 3.1: Theorems III.2 and III.3 present bounds on the out-of-sample prediction errors of the solutions to (4) and (5), respectively. The bounds depend on the average training loss and the magnitude of the regularizer. It can be concluded that using a regularized learning procedure improves the prediction accuracy.

IV. Experiments

In this section we will test the two relaxations (4) and (5) on a number of synthetic datasets, and compare them against several other popular methods for MLR, including OLS, Reduced Rank Regression (RRR) [19, 8], Principal Components Regression (PCR) [9], FES [7], the Curds and Whey (C&W) procedure [5], and Ridge Regression (RR) [20, 10].

To test the robustness of various methods, we inject outliers to the datasets whose distribution differs from the majority by a normally distributed random quantity. Note that the perturbation occurs only on the response variables. The data generation process can be described as follows.



(a) $\sigma^{\mathbf{y}}_{ij} = 0.5^{|i-j|}$. (b) $\sigma^{\mathbf{y}}_{ij} = 0.1^{|i-j|}$.

Fig. 1. The out-of-sample WMSE as r varies.

1) Generate each element of the true coefficient matrix $\mathbf{B}^* \in \mathbb{R}^{p\times K}$ with $p = 8, K = 3$ from the standard normal distribution.
2) Generate $\mathbf{x} \in \mathbb{R}^p$ from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_x})$, where $\mathbf{\Sigma_x} = (\sigma^{\mathbf{x}}_{ij})^{j=1,\ldots,p}_{i=1,\ldots,p}$ has ones on the diagonal, and off-diagonal elements specified by $\sigma^{\mathbf{x}}_{ij} = 0.7^{|i-j|},\ i \neq j$.
3) For a clean sample, generate $\mathbf{y}$ from $\mathcal{N}((\mathbf{B}^*)'\mathbf{x}, \mathbf{I}_K)$; for outliers, generate $\mathbf{y}$ from $\mathcal{N}((\mathbf{B}^*)'\mathbf{x}, \mathbf{I}_K) + \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_y})$, where $\mathbf{\Sigma_y} = (\sigma^{\mathbf{y}}_{ij})^{j=1,\ldots,K}_{i=1,\ldots,K}$ has ones on the diagonal, and off-diagonal elements specified by $\sigma^{\mathbf{y}}_{ij} = 0.5^{|i-j|},\ i \neq j$.

We generate $N = 60$ training samples that contain a proportion r of outliers to train the MLR models mentioned above, and compare their performance on a set of $M = 40$ test samples. Note that the test samples do not contain any outlier, as we expect the estimated regression coefficients to be consistent with the clean data distribution (robust to outliers), and only care about their predictive performance on clean data samples.

All parameters, including the regularization coefficients of our methods and RR, the number of principal components used in PCR, the optimal rank in RRR, are tuned using cross-validation. We evaluate the following *Weighted Mean Squared Error (WMSE)* on the test set:

$$\text{WMSE} \triangleq \frac{1}{M}\sum_{i=1}^{M}(\mathbf{y}_i-\hat{\mathbf{y}}_i)'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\mathbf{y}}_i),$$

where $\mathbf{y}_i, \hat{\mathbf{y}}_i$ are the true and predicted responses for the i-th test sample, and $\hat{\mathbf{\Sigma}} = (\mathbf{Y}-\hat{\mathbf{Y}})'(\mathbf{Y}-\hat{\mathbf{Y}})/(N-pK)$, with $\mathbf{Y}, \hat{\mathbf{Y}} \in \mathbb{R}^{N\times K}$ the true and estimated response matrices on the training set, respectively.

We plot the simulation results in Fig. 1 as the proportion of outliers r is varied. Note that we rename relaxation (4) as DRO-Decomp since it can be decomposed into K independent sub-problems. DRO-MLR corresponds to formulation (5).

To investigate the effect of the noise covariance $\mathbf{\Sigma_y}$ on the performance, we decrease the value of $\sigma^{\mathbf{y}}_{ij}$ to $0.1^{|i-j|}$. By comparing the two figures, we conclude:

1) Both DRO-MLR and DRO-Decomp achieve a smaller prediction bias than other methods, with DRO-MLR slightly better than DRO-Decomp.
2) PCR has a similar performance to our methods when the proportion of outliers r is low. As r increases,

however, our methods perform better.

3) When the added noise becomes less correlated, PCR achieves a comparable performance to us. Our methods are more advantageous when the response variables are highly correlated.

The success of PCR is due to the fact that it eliminates the multicollinearity through transforming the original predictors into a set of uncorrelated Principal Components (PCs). It can result in dimension reduction through excluding some of the low variance PCs. Essentially, PCR transforms the problem to one that is decomposable. However, PCR is criticized for not offering easily interpretable models due to the linear transformation of the predictors. By contrast, our methods retain the structure of the predictors, yielding a model that has a comparable performance to PCR without sacrificing interpretability.

V. Conclusions

We proposed a *Distributionally Robust Optimization (DRO)* formulation for *Multivariate Linear Regression (MLR)* to estimate a robust regression coefficient matrix that is immunized against large noise in the data. A regularized regression reformulation was derived using a newly defined matrix norm that scalarizes each column by the sum of the absolute values of its elements. We provided bounds on its prediction bias, and empirically tested its performance on a number of synthetic datasets, showing that our approach results in a smaller prediction error compared to a series of alternatives.

References

[1] H. Zhang, H. Zhao, J. Sun, D. Wang, and K. Kim, "Regression analysis of multivariate panel count data with an informative observation process," *Journal of Multivariate Analysis*, vol. 119, pp. 71–80, 2013.

[2] B. Hidalgo and M. Goodman, "Multivariate or multivariable regression?" *American journal of public health*, vol. 103, no. 1, pp. 39–40, 2013.

[3] R. S. Tsay, *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2013.

[4] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human brain mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[5] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 1, pp. 3–54, 1997.

[6] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 615–637, 2005.

[7] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 329–346, 2007.

[8] R. Velu and G. C. Reinsel, *Multivariate reduced-rank regression: theory and applications*. Springer Science & Business Media, 2013, vol. 136.

[9] W. F. Massy, "Principal components regression in exploratory statistical research," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 234–256, 1965.

[10] Y. Haitovsky, "On multivariate ridge regression," *Biometrika*, vol. 74, no. 3, pp. 563–570, 1987.

[11] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[12] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.

[13] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 517–564, 2018.

[14] S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," in *Advances in Neural Information Processing Systems*, 2015, pp. 1576–1584.

[15] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *arXiv preprint arXiv:1710.10016*, 2017.

[16] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," *arXiv preprint arXiv:1712.06050*, 2017.

[17] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured schatten norm regularization," in *Advances in neural information processing systems*, 2013, pp. 1331–1339.

[18] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[19] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of multivariate analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[20] P. J. Brown, J. V. Zidek *et al.*, "Adaptive multivariate ridge regression," *The Annals of Statistics*, vol. 8, no. 1, pp. 64–74, 1980.