JOTA manuscript No.
(will be inserted by the editor)

Robust Grouped Variable Selection Using Distributionally Robust Optimization

Ruidi Chen · Ioannis Ch. Paschalidis

Received: date / Accepted: date

Abstract We propose a *Distributionally Robust Optimization (DRO)* formulation with a Wasserstein-based uncertainty set for selecting grouped variables under perturbations on the data for both linear regression and classification problems. The resulting model offers robustness explanations for *Grouped Least Absolute Shrinkage and Selection Operator (GLASSO)* algorithms and highlights the connection between robustness and regularization. We prove probabilistic bounds on the out-of-sample loss and the estimation bias, and establish the grouping effect of our estimator, showing that coefficients in the same group converge to the same value as the sample correlation between covariates approaches 1. Based on this result, we propose to use the spectral clustering algorithm with the Gaussian similarity function to perform grouping on the predictors, which makes our approach applicable without knowing the grouping structure a priori. We compare our approach to an array of alternatives and provide extensive numerical results on both synthetic data and a real large dataset of surgery-related medical records, showing that our formulation produces an interpretable and parsimonious model that encourages sparsity at a group level and is able to achieve better prediction and estimation performance in the presence of outliers.

Keywords Regression · Classification · Grouped LASSO · Wasserstein Metric · Spectral Clustering

Mathematics Subject Classification (2000) 90C47 · 62J07 · 62J12

Ruidi Chen, Boston University, Boston, MA, USA, rchen15@bu.edu.

· Ioannis Ch. Paschalidis, Corresponding author, Boston University, Boston, MA, USA, yannisp@bu.edu.

1 Introduction

We consider the problem of finding a robust regression/classification plane under perturbations on the training data, when there exists a predefined grouping structure for the predictors, e.g., encoding a categorical predictor using a group of indicator variables. The goal is to jointly select/drop all variables in a group, i.e., induce *group sparsity*, and produce robust estimates that generalize well out of sample. Grouped variable selection gives rise to more interpretable models. Moreover, group sparsity leads to an estimation error of regression coefficients that scales with the number of groups and group sizes, instead of with the raw number of features in the regression model [17,20].

To perform variable selection at a group level, the *Grouped Least Absolute Shrinkage and Selection Operator (GLASSO)* was proposed by [1,33]. Several extensions have been explored in later works, see [34, 18,28,7]. The group sparsity in general regression/classification models has also been investigated, see, for example, [21] for GLASSO in logistic regression, and [24] for GLASSO in generalized linear models. We note that most of the existing works endeavor to generalize/modify the GLASSO formulation heuristically to achieve various goals. However, few of those works were able to provide a rigorous explanation or theoretical justification for the form of the penalty term.

In this work we attempt to fill this gap by casting the robust grouped variable selection problem into a *Distributionally Robust Optimization (DRO)* framework, which induces robustness via minimizing a worst-case expected loss function over a probabilistic ambiguity set that is constructed from the observed samples and characterized by certain known properties of the true data-generating distribution. DRO has been an active area of research in recent years, due to its probabilistic interpretation of the uncertain data, tractability when assembled with certain metrics, and extraordinary performance observed on numerical examples, see, for example, [14,13,26,12,8,9]. [11] demonstrated the advantage of DRO through providing its finite sample and asymptotic convergence properties, showing that DRO often exhibits improved generalization and tail performance. The uncertainty set in DRO can be constructed (*i*) through a moment ambiguity set [10,15,36], or (*ii*) as a ball of distributions centered at some nominal distribution defined via some probabilistic distance metric such as the ϕ-divergence, the Prokhorov metric, and the Wasserstein distance.

We consider a DRO formulation with the uncertainty set being a ball of distributions defined via the Wasserstein metric, motivated by the fact that (i) the Wasserstein metric takes into account the closeness between support points while other metrics only consider the probabilities on these points, and (ii) the Wasserstein ambiguity set is rich enough to contain both continuous and discrete relevant distributions, while other metrics such as the Kullback-Leibler (KL) divergence, do not allow for probability mass outside the support of the nominal distribution. We show that in *Least Absolute Deviation (LAD)* and *logistic regression (LG)*, for both non-overlapping and overlapping predictor groups, the Wasserstein DRO model can be reformulated as a regularized empirical loss minimization problem, where the regularizer coincides with the GLASSO penalty, and its magnitude is equal to the radius of the distributional ambiguity set. Through such a reformulation we establish a connection between regularization and robustness and offer new insights into the GLASSO penalty term.

We should note that such a connection between robustification and regularization under norm-bounded deterministic disturbances in the predictors has been discovered in [31,32,3]. Within the Wasserstein DRO framework, such an equivalence has been established for LG in [25], and for LAD regression in [8]. More recently, [26,13] have provided a unified framework for connecting the Wasserstein DRO with regularized learning procedures. None of the aforementioned works, however, considered grouped variable selection; our work sheds new light on the significance of exploring the group-wise DRO problem. It is worth noting that [5] has studied the group-wise regularization estimator with the square root of the expected loss under the Wasserstein DRO framework and recovered the *Grouped Square Root LASSO (GSRL)*. Here, we present a more general framework that includes both the LAD and the negative log-likelihood loss functions, under both non-overlapping and overlapping group structures. Moreover, we point out the potential of generalizing such results to a class of loss functions with a finite growth rate.

Another contribution of this work lies in adding a correlation-based pre-clustering step to GLASSO, as a consequence of a grouping effect result derived specifically for our DRO GLASSO estimator. This has a similar flavor to [6], where they considered a pre-clustering step based on either the canonical correlation between groups or the sample correlation between covariates and validated their approach from the standpoint of statistical consistency. Here, we justify the correlation-based clustering from the optimization point of view, by analyzing the optimality conditions satisfied by the DRO GLASSO estimator. We summarize our contributions as follows.

- We propose a Wasserstein DRO formulation for inducing group sparsity under perturbations on the data for linear regression and classification with both non-overlapping and overlapping predictor groups.
- We establish a connection between robustness and regularization under group sparsity, offering robustness explanations for GLASSO algorithms.
- We show probabilistic bounds on the prediction and estimation errors of the DRO estimator, and establish its grouping effect from an optimization perspective.
- We propose a purely data-driven, correlation-based pre-clustering step to DRO GLASSO, rendering our model applicable when the group structure is not known a priori.
- We validate the superiority of our DRO GLASSO model through providing extensive numerical results on both synthetic data and a large dataset of surgery-related medical records.

The remainder of the paper is organized as follows. Section 2 introduces the Wasserstein GLASSO formulations for LAD and LG. Section 3 establishes a desirable grouping effect for the solutions, which leads to a correlation-based pre-clustering step on the predictors. Section 4 presents numerical results on both synthetic data and a real very large dataset with surgery-related medical records. Conclusions are in Section 5. Proofs and additional numerical results are in the Appendix.

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 $\mathbb{P}$ the probability of an event. All vectors are column vectors. For space saving reasons, we write $\mathbf{x} = (x_1, \ldots, x_n)$ to denote the column vector $\mathbf{x} \in \mathbb{R}^n$. We use prime to denote transpose, $\|\cdot\|$ for the general norm operator, and $\|\mathbf{x}\|_p \triangleq (\sum_i |x_i|^p)^{1/p}$ for the ℓ_p norm, where $p \geq 1$.

2 Problem Formulation

In this section we describe the model setup and derive what we call the *Groupwise Wasserstein Grouped LASSO (GWGL)* formulation for a LAD regression model and an LG model.

2.1 GWGL for Continuous Response Variables

Consider a linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_N)$ is the response vector, $\mathbf{X}$ is an $N \times p$ design matrix, with i-th row $\mathbf{x}_i'$ being the predictor vector for the i-th sample, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the vector of regression coefficients, and $\boldsymbol{\eta} \in \mathbb{R}^N$ is a random noise vector. We assume that the predictors belong to L prescribed groups, with group size p_l, $l = 1, \ldots, L$, and $\sum_{l=1}^{L} p_l = p$ (no overlap among groups). We use $\mathbf{x}_{,j} \in \mathbb{R}^N$ to denote the j-th column of $\mathbf{X}$, corresponding to the j-th predictor. A p_l-dimensional vector $\boldsymbol{\beta}^l$ denotes the vector of regression coefficients for group l. For a generic predictor vector $\mathbf{x} \in \mathbb{R}^p$, we decompose it into L groups $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, each $\mathbf{x}^l$ containing the p_l predictors of group l.

The main assumption we make regarding $\boldsymbol{\beta}^*$ is that it is *group sparse*, i.e., $\boldsymbol{\beta}^l = \mathbf{0}$ for l in some subset of $\{1, \ldots, L\}$. Our goal is to obtain an accurate estimate of $\boldsymbol{\beta}^*$ under perturbations on $(\mathbf{X}, \mathbf{y})$. Suppose we have N i.i.d. samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$. We model stochastic disturbances on the data via distributional uncertainty, and apply a Wasserstein DRO framework to inject robustness into the solution. Our learning problem is formulated as:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}\big[|y - \mathbf{x}'\boldsymbol{\beta}|\big], \tag{2}$$

where $(\mathbf{x}, y) \in \mathbb{R}^{p+1}$ denotes a generic predictor-response pair; and $\mathbb{Q}$ is the probability distribution of $(\mathbf{x}, y)$. The inner optimization problem is over $\mathbb{Q}$ in some set Ω defined as:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \le \epsilon\}, \tag{3}$$

where $\epsilon > 0$ specifies the size of the ambiguity set Ω, $\mathcal{Z}$ is the set of possible values for $(\mathbf{x}, y)$, $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$, $\hat{\mathbb{P}}_N$ is the empirical probability distribution that assigns equal probability on each training sample point $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, and $W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the order-one Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ defined on the metric space $(\mathcal{Z}, s)$ by:

$$W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \triangleq \min_{\Pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} \left\{ \int_{\mathcal{Z} \times \mathcal{Z}} s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\ \Pi\big(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)\big) \right\}, \tag{4}$$

where we use the metric $s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) = \|(\mathbf{x}_1, y_1) - (\mathbf{x}_2, y_2)\|$ for the regression setting; and Π is the joint distribution of $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, respectively.

We assume that all the N training samples $(\mathbf{x}_i, y_i), i = 1, \ldots, N$, are independent and identical realizations of $(\mathbf{x}, y)$, which comes from a mixture of two distributions, with probability q from an "outlying" distribution $\mathbb{P}_{\text{out}}$ and with probability $1-q$ from the true distribution $\mathbb{P}$. Our goal is to generate estimators that are consistent with the true distribution $\mathbb{P}$. We next show that if $q < 0.5$, and ϵ chosen judiciously, this is possible.

Theorem 2.1 *Suppose we are given two probability distributions $\mathbb{P}$ and $\mathbb{P}_{out}$, and the mixture distribution $\mathbb{P}_{mix}$ is a convex combination of the two: $\mathbb{P}_{mix} = q\mathbb{P}_{out} + (1-q)\mathbb{P}$. Then,*

$$\frac{W_1(\mathbb{P}_{out}, \mathbb{P}_{mix})}{W_1(\mathbb{P}, \mathbb{P}_{mix})} = \frac{1-q}{q}.$$

Theorem 2.1 implies that when $q < 0.5$, and $W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) \le \epsilon < W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})$, for a large enough sample size (so that $\hat{\mathbb{P}}_N$ is a good approximation of $\mathbb{P}_{\text{mix}}$), the probabilistic ambiguity set Ω will include the true distribution and exclude the outlying one, thus providing protection against the disturbances.

The formulation in (2) is robust since it minimizes over the regression coefficients the worst case expected loss; the latter being the expected loss maximized over all probability distributions in the ambiguity set Ω. Formulation (2) injects additional robustness by adopting the LAD loss, rendering it more robust to large residuals and yielding a smaller estimation bias [8].

It has been shown in [8] that (2) could be relaxed to:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon \|(-\boldsymbol{\beta}, 1)\|_*, \tag{5}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, which is the norm used to define the distance function s in the Wasserstein metric (4). The dual norm is defined as $\|\boldsymbol{\theta}\|_* \triangleq \sup_{\|\mathbf{z}\| \le 1} \boldsymbol{\theta}'\mathbf{z}$. Our GWGL formulation will be derived as a special case of (5), using a specific notion of norm on the $(\mathbf{x}, y)$ space that reflects the group structure of the predictors and takes into account the group sparsity requirement. Specifically, for a vector $\mathbf{z}$ with a group structure $\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L)$, define its (q, t)-norm, with $q, t \ge 1$, as:

$$\|\mathbf{z}\|_{q,t} = \Big(\sum_{l=1}^{L} \big(\|\mathbf{z}^l\|_q \big)^t \Big)^{1/t}.$$

The (q, t)-norm of $\mathbf{z}$ is actually the ℓ_t-norm of the vector $(\|\mathbf{z}^1\|_q, \ldots, \|\mathbf{z}^L\|_q)$, which represents each group vector $\mathbf{z}^l$ in a concise way via the ℓ_q-norm.

Inspired by the LASSO where the ℓ_1-regularizer is used to induce sparsity on the individual level, we wish to deduce an ℓ_1-norm penalty from (5) on the group level to induce group sparsity on $\boldsymbol{\beta}^*$. This motivates the use of the $(2,\infty)$-norm on the weighted predictor-response vector $\mathbf{z_w} \triangleq (\frac{1}{\sqrt{p_1}}\mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}}\mathbf{x}^L, My)$, where the weight vector is $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M)$, and M is a positive weight assigned to the response. Specifically,

$$\|\mathbf{z_w}\|_{2,\infty} = \max\left\{\frac{1}{\sqrt{p_1}}\|\mathbf{x}^1\|_2, \ldots, \frac{1}{\sqrt{p_L}}\|\mathbf{x}^L\|_2, M|y|\right\}. \tag{6}$$

In (6) we normalize each group by the number of predictors, to prevent large groups from having a large impact on the distance metric. The $\|\cdot\|_{2,\infty}$ operator computes the maximum of the ℓ_2 norms of the (weighted) grouped predictors and the response. It essentially selects the most influential group when determining the closeness between two points in the predictor-response space, which is consistent with our group sparsity assumption in that not all groups of predictors contribute to the determination of y, and thus a metric that ignores the unimportant groups (e.g., $\|\cdot\|_{2,\infty}$) is desired.

To obtain the GWGL formulation, we need to derive the dual norm of $\|\cdot\|_{2,\infty}$. A general result that applies to any (q,t)-norm is presented in the following theorem.

Theorem 2.2 *Consider a vector* $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$*, where each* $\mathbf{x}^l \in \mathbb{R}^{p_l}$*, and* $\sum_l p_l = p$*. Define the weighted* (r,s)*-norm of* $\mathbf{x}$ *with the weight vector* $\mathbf{w} = (w_1, \ldots, w_L)$ *to be:*

$$\|\mathbf{x_w}\|_{r,s} = \left(\sum_{l=1}^{L}\left(\|w_l\mathbf{x}^l\|_r\right)^s\right)^{1/s},$$

where $\mathbf{x_w} = (w_1\mathbf{x}^1, \ldots, w_L\mathbf{x}^L)$*,* $w_l > 0, \forall l$*, and* $r, s \geq 1$*. Then, the dual norm of the weighted* (r,s)*-norm with weight* $\mathbf{w}$ *is the* (q,t)*-norm with weight* $\mathbf{w}^{-1}$*, where* $1/r + 1/q = 1$*,* $1/s + 1/t = 1$*, and* $\mathbf{w}^{-1} = (1/w_1, \ldots, 1/w_L)$*.*

Now, let us go back to (6), which is the weighted $(2,\infty)$-norm of $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^L, y)$ with the weight $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M)$. According to Theorem 2.2, the dual norm of the weighted $(2,\infty)$-norm with weight $\mathbf{w}$ evaluated at some $\tilde{\boldsymbol{\beta}} = (-\boldsymbol{\beta}^1, \ldots, -\boldsymbol{\beta}^L, 1)$ is:

$$\|\tilde{\boldsymbol{\beta}}_{\mathbf{w}^{-1}}\|_{2,1} = \sum_{l=1}^{L}\sqrt{p_l}\|\boldsymbol{\beta}^l\|_2 + \frac{1}{M},$$

where $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L}, 1/M)$. Therefore, the GWGL formulation for Linear Regression (GWGL-LR) takes the following form:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\sum_{l=1}^{L}\sqrt{p_l}\|\boldsymbol{\beta}^l\|_2, \tag{7}$$

where the constant term $1/M$ has been removed. We see that by using the weighted $(2,\infty)$-norm in the predictor-response space, we are able to recover the commonly used penalty term for GLASSO [1,33]. Our

Wasserstein DRO framework offers new interpretations for the GLASSO penalty from the standpoint of the distance metric on the predictor-response space and establishes the connection between group sparsity and distributional robustness.

2.2 GWGL for Binary Categorical Response Variables

In this subsection we will explore the GWGL formulation for binary classification problems. Let $\mathbf{x} \in \mathbb{R}^p$ denote the predictor and $y \in \{-1, +1\}$ the associated binary label to be predicted. In LG, the conditional distribution of y given $\mathbf{x}$ is modeled as

$$\mathbb{P}(y|\mathbf{x}) = \big(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})\big)^{-1},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown coefficient vector (classifier) to be estimated. The *Maximum Likelihood Estimator (MLE)* of $\boldsymbol{\beta}$ is found by minimizing the *negative log-likelihood (logloss)*:

$$l_{\boldsymbol{\beta}}(\mathbf{x}, y) = \log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})).$$

To apply the Wasserstein DRO framework, we define the following distance metric:

$$s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \triangleq \begin{cases} \|\mathbf{x}_1 - \mathbf{x}_2\|, & \text{if } y_1 = y_2, \\ \infty, & \text{otherwise.} \end{cases} \tag{8}$$

Through (8) we emphasize the role of y in determining the distance between data points, i.e., samples from different classes are considered to be infinitely far away from each other. Our robust LG problem is modeled as:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}\big[\log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x}))\big], \tag{9}$$

where $\mathbb{Q}$ is the probability distribution of $(\mathbf{x}, y)$, belonging to some set Ω that includes all probability distributions whose order-one Wasserstein distance (on the metric space $(\mathcal{Z}, s)$ where $\mathcal{Z} = \mathbb{R}^p \times \{-1, +1\}$) to the empirical distribution $\hat{\mathbb{P}}_N$ is no more than ϵ. In the following theorem, we reformulate (9).

Theorem 2.3 *Suppose we observe N realizations of the data, denoted by $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$. When the Wasserstein metric is induced by (8), the DRO problem (9) can be reformulated as:*

$$\inf_{\boldsymbol{\beta}} \mathbb{E}^{\hat{\mathbb{P}}_N}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] + \epsilon\|\boldsymbol{\beta}\|_* = \inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \log\big(1 + \exp(-y_i\boldsymbol{\beta}'\mathbf{x}_i)\big) + \epsilon\|\boldsymbol{\beta}\|_*. \tag{10}$$

We note that [25,26,13] arrive at a similar formulation to (10) by other means of derivation. Different from these existing works, we will consider specifically the application of (10) to grouped predictors where the goal is to induce group level sparsity on the coefficients/classifier. As in Section 2.1, we assume that the predictor vector $\mathbf{x}$ can be decomposed into L groups, i.e., $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, each $\mathbf{x}^l$ containing p_l predictors of group l, and $\sum_{l=1}^{L} p_l = p$ (no overlap among groups). To reflect the group sparse structure, we consider the $(2, \infty)$-norm of the weighted predictor vector $\mathbf{x}_\mathbf{w} \triangleq (\frac{1}{\sqrt{p_1}}\mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}}\mathbf{x}^L)$, where the weight vector is $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}})$. According to Theorem 2.2, the dual norm of the weighted $(2, \infty)$-norm with weight $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}})$ evaluated at $\boldsymbol{\beta}$ is:

$$\|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} = \sum_{l=1}^{L} \sqrt{p_l}\|\boldsymbol{\beta}^l\|_2,$$

where $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L})$, and $\boldsymbol{\beta}^l$ denotes the vector of coefficients corresponding to group l. Therefore, the GWGL formulation for LG (GWGL-LG) takes the form:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \log\big(1 + \exp(-y_i\boldsymbol{\beta}'\mathbf{x}_i)\big) + \epsilon\sum_{l=1}^{L} \sqrt{p_l}\|\boldsymbol{\beta}^l\|_2. \tag{11}$$

The above derivation techniques also apply to other loss functions whose growth rate is finite, e.g., the hinge loss used by the *Support Vector Machine (SVM)*, and therefore, the GWGL SVM model can be developed in a similar fashion. It is also worth noting that the regularizer in our tractable reformulation (10) is related to the growth rate of the loss function, with the magnitude of the penalty being the radius of the Wasserstein ball [8,13]. This enables new perspectives of the regularization term and provides guidance on the selection/tuning of the regularization coefficient.

2.3 GLASSO with Overlapping Groups

In this subsection we will explore the GLASSO formulation with overlapping groups, and show that our Wasserstein DRO framework recovers a latent GLASSO approach that was first proposed in [23].

When the groups overlap with each other, the penalty term $\sum_{l=1}^{L} \sqrt{p_l}\|\boldsymbol{\beta}^l\|_2$ leads to a solution whose support is almost surely the complement of a union of groups [19]. In other words, setting one group to zero shrinks its covariates to zero even if they belong to other groups, in which case these other groups will not be entirely selected. [23] proposed a latent GLASSO approach where they introduce a set of latent variables that induce a solution vector whose support is a union of groups, so that the estimator would select entire groups of covariates. Specifically, define the latent variables $\mathbf{v}^l \in \mathbb{R}^p,\ l = 1, \ldots, L$, such that $\text{supp}(\mathbf{v}^l) \subset \mathcal{G}^l,\ l = 1, \ldots, L$, where $\text{supp}(\mathbf{v}^l) \subset \{1, \ldots, p\}$ denotes the support of $\mathbf{v}^l$, i.e., the set of predictors $i \in \{1, \ldots, p\}$ such that $v_i^l \neq 0$, and $\mathcal{G}^l$ denotes the set of predictors that are in group l. Our assumption is that $\exists\ l_1, l_2$ such that $\mathcal{G}^{l_1} \cap \mathcal{G}^{l_2} \neq \emptyset$. The latent GLASSO formulation has the form:

$$\begin{aligned} \inf_{\boldsymbol{\beta}, \mathbf{v}^1, \ldots, \mathbf{v}^L} \quad & \frac{1}{N} \sum_{i=1}^N l_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) + \epsilon \sum_{l=1}^L d_l \|\mathbf{v}^l\|_2, \\ \text{s.t.} \quad & \boldsymbol{\beta} = \sum_{l=1}^L \mathbf{v}^l, \end{aligned} \tag{12}$$

where $l_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i)$ denotes the loss at sample $(\mathbf{x}_i, y_i)$, and d_l is a user-specified penalty strength of group l. Let $\hat{\mathbf{v}}^l$, $l = 1, \ldots, L$, denote an optimal solution of (12). By using the latent vectors, Formulation (12) has the flexibility of implicitly adjusting the support of the latent vectors such that for any $i \in \text{supp}(\hat{\mathbf{v}}^l)$ where $\hat{\mathbf{v}}^l = \mathbf{0}$, it does not belong to the support of any non-shrunk latent vectors. As a result, the covariates that belong to both shrunk and non-shrunk groups would not be mistakenly driven to zero. Formulation (12) favors solutions which shrink some $\mathbf{v}^l$ to zero, while the non-shrunk components satisfy $\text{supp}(\mathbf{v}^l) = \mathcal{G}^l$, therefore leading to estimators whose support is the union of a set of groups.

To show that (12) can be obtained from the Wasserstein DRO framework, we consider the following weighted $(2, \infty)$-norm on the predictor space:

$$s(\mathbf{x}) = \max_l d_l^{-1} \|\mathbf{x}^l\|_2. \tag{13}$$

For simplicity, we treat the response y as a deterministic quantity so that the Wasserstein metric is defined only on the predictor space. The scenario with stochastic responses can be accommodated by introducing some constant M. [23] showed that the dual norm of (13) is $\Theta(\boldsymbol{\beta}) \triangleq \sum_{l=1}^L d_l \|\mathbf{v}^l\|_2$, with $\boldsymbol{\beta} = \sum_{l=1}^L \mathbf{v}^l$, and $\boldsymbol{\beta} \mapsto \Theta(\boldsymbol{\beta})$ is a valid norm. By reformulating (12) as:

$$\inf_{\boldsymbol{\beta}} \quad \frac{1}{N} \sum_{i=1}^N l_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) + \epsilon \Theta(\boldsymbol{\beta}), \text{ with } \Theta(\boldsymbol{\beta}) = \min_{\substack{\mathbf{v}^1, \ldots, \mathbf{v}^L, \\ \sum_{l=1}^L \mathbf{v}^l = \boldsymbol{\beta}}} \sum_{l=1}^L d_l \|\mathbf{v}^l\|_2, \tag{14}$$

we have shown that (12) can be derived as a consequence of the Wasserstein DRO formulation with the Wasserstein metric induced by (13). In fact, (14) is equivalent to a regular GLASSO in a covariate space of higher dimension obtained by duplication of the covariates belonging to several groups. For simplicity our subsequent analysis assumes non-overlapping groups.

3 Grouping Effect of the Estimators

In this section we establish a *grouping effect* for the solutions to GWGL-LR and GWGL-LG, which measures the similarity of the estimated coefficients in the same group. Ideally, for highly correlated predictors in the same group, it is desired that their coefficients are close so that they can be jointly selected/dropped (group sparsity). The discussion on the prediction and estimation quality of the solutions is deferred to Appendix A.

To investigate the grouping effect of the estimators, we examine the difference between coefficient estimates as a function of the sample correlation between their corresponding predictors in the following theorem.

Theorem 3.1 *Suppose the predictors are standardized (columns of* $\mathbf{X}$ *have zero mean and unit variance). Let* $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ *be the optimal solution to (7) (or (11)). If* $\mathbf{x}_{,i}$ *is in group* l_1 *and* $\mathbf{x}_{,j}$ *is in group* l_2*, and* $\|\hat{\boldsymbol{\beta}}^{l_1}\|_2 \neq 0$*,* $\|\hat{\boldsymbol{\beta}}^{l_2}\|_2 \neq 0$*, define*

$$D(i,j) = \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right|.$$

Then,

$$D(i,j) \leq \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon},$$

where $\rho = \mathbf{x}_{,i}'\mathbf{x}_{,j}$ *is the sample correlation, and* p_{l_1}, p_{l_2} *are the number of predictors in groups* l_1 *and* l_2*, respectively.*

Theorem 3.1 establishes a unified result for the grouping effect of the GWGL-LR and GWGL-LG solutions. When $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ are both in group l and $\|\hat{\boldsymbol{\beta}}^{l}\|_2 \neq 0$, it follows

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{\sqrt{2(1-\rho)}\|\hat{\boldsymbol{\beta}}^{l}\|_2}{\epsilon\sqrt{N p_l}}. \tag{15}$$

From (15) we see that as the within group correlation increases, the difference between $\hat{\beta}_i$ and $\hat{\beta}_j$ becomes smaller. In the extreme case where $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ are perfectly correlated, $\hat{\beta}_i = \hat{\beta}_j$. This grouping effect enables recovery of sparsity on a group level when the correlation between predictors in the same group is high, and implies the use of predictors' correlation as a grouping criterion. One of the popular clustering algorithms, called *spectral clustering* [27,22], performs grouping based on the eigenvalues/eigenvectors of the Laplacian matrix of the similarity graph that is constructed using the *similarity matrix* of data, and divides the data points (predictors) into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. The similarity matrix measures the pairwise similarities between data points, which in our case could be the pairwise correlations between predictors.

4 Numerical Results

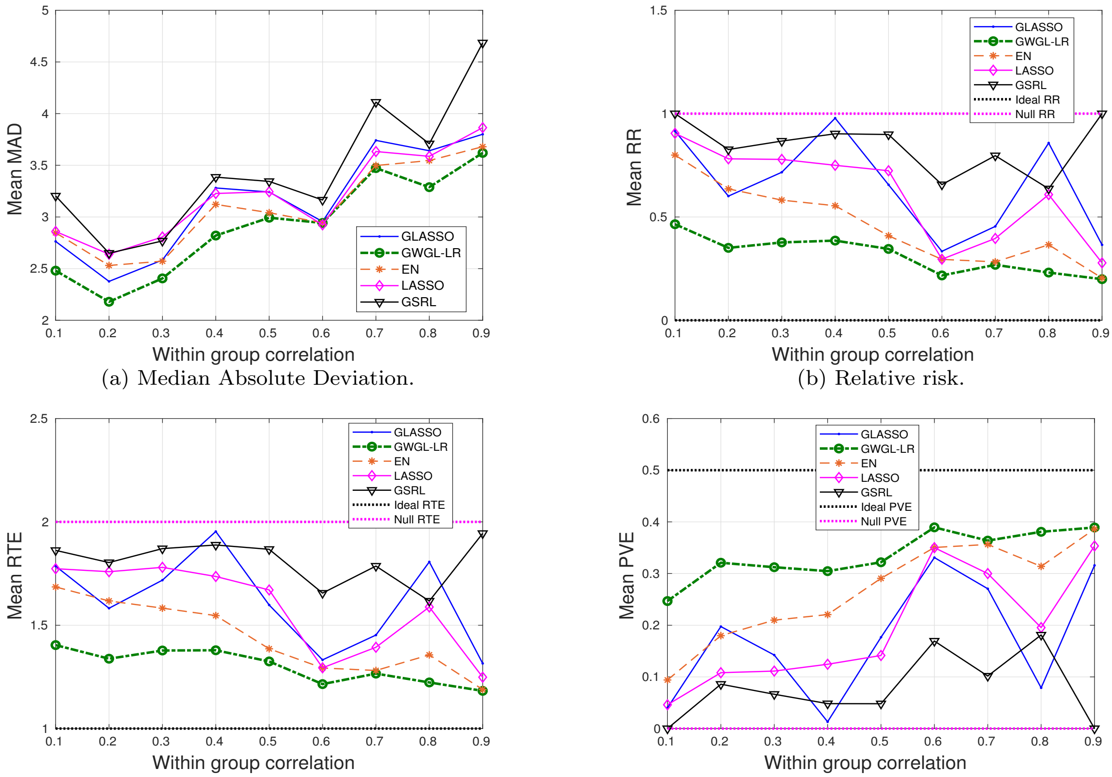
In this section we compare our GWGL formulations with other commonly used predictive models. In the linear regression setting, we compare GWGL-LR with models that either (*i*) use a different loss function, e.g., the traditional GLASSO with an ℓ_2-loss [33], and the Group Square-Root LASSO (GSRL) [7] that minimizes the square root of the ℓ_2-loss; or (*ii*) do not make use of the grouping structure of the predictors, e.g., the Elastic Net (EN) [35], and the LASSO [29]. For classification problems, we consider alternatives that minimize the empirical logloss plus penalty terms that do not utilize the grouping structure of the predictors, e.g., the ℓ_1-regularizer (LG-LASSO), ℓ_2-regularizer (LG-Ridge), and their combination (LG-EN).

4.1 GWGL-LR on Synthetic Datasets

In this subsection we will compare GWGL-LR with the aforementioned models on several synthetic datasets. The data are generated as follows. (1) Set β_i^* to 0.5 if predictor i belongs to an even group, and 0 otherwise. (2) Generate $\mathbf{x} \in \mathbb{R}^p$ from the Gaussian distribution $\mathcal{N}_p(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{i,j})_{i,j=1}^p$ has diagonal elements equal to 1, and off-diagonal elements $\sigma_{i,j}$ equal to ρ_w if predictors i and j are in the same group, and 0 otherwise. Here, ρ_w is called the *within group correlation.* (3) With probability $1 - q$, generate y from $\mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$, and with probability q, generate y from $\mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^* + 5\sigma, \sigma^2)$, where σ^2 is the intrinsic variance of y, and q is the probability of abnormal samples (outliers).

We generate 10 datasets consisting of $N = 100$ training samples and $M_t = 60$ test samples with 4 groups of predictors, where $p_1 = 1, p_2 = 3, p_3 = 5, p_4 = 7$, and $p = \sum_{i=1}^4 p_i = 16$. We are interested in studying the impact of (*i*) *Signal to Noise Ratio (SNR)*, defined as: $\text{SNR} = (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*/\sigma^2$, and (*ii*) the *within group correlation* ρ_w. The performance metrics are: (*i*) *Median Absolute Deviation (MAD)* on the test set, defined as the median value of $|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}|,\ i = 1, \ldots, M_t$, with $\hat{\boldsymbol{\beta}}$ being the estimate of $\boldsymbol{\beta}^*$ obtained from the training set, and $(\mathbf{x}_i, y_i),\ i = 1, \ldots, M_t$, being the observations from the test set; (*ii*) *Relative Risk (RR)*, *Relative Test Error (RTE)*, and *Proportion of Variance Explained (PVE)* of $\hat{\boldsymbol{\beta}}$ (definitions in Appendix B).

Before solving for the regression coefficients, the grouping of predictors needs to be determined. Unlike most of the existing works where the grouping structure is assumed to be known a priori, we propose to use a data-driven clustering algorithm to group the predictors based on their sample correlations. Specifically, we consider the *spectral clustering* [27,22] algorithm with the Gaussian similarity function $\text{Gs}(\mathbf{x}_{,i}, \mathbf{x}_{,j}) \triangleq \exp\big(-\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2/(2\sigma_s^2)\big)$ that captures the sample pairwise correlations between predictors, where σ_s is some scale parameter whose selection is discussed in Appendix B.



(a) Median Absolute Deviation.

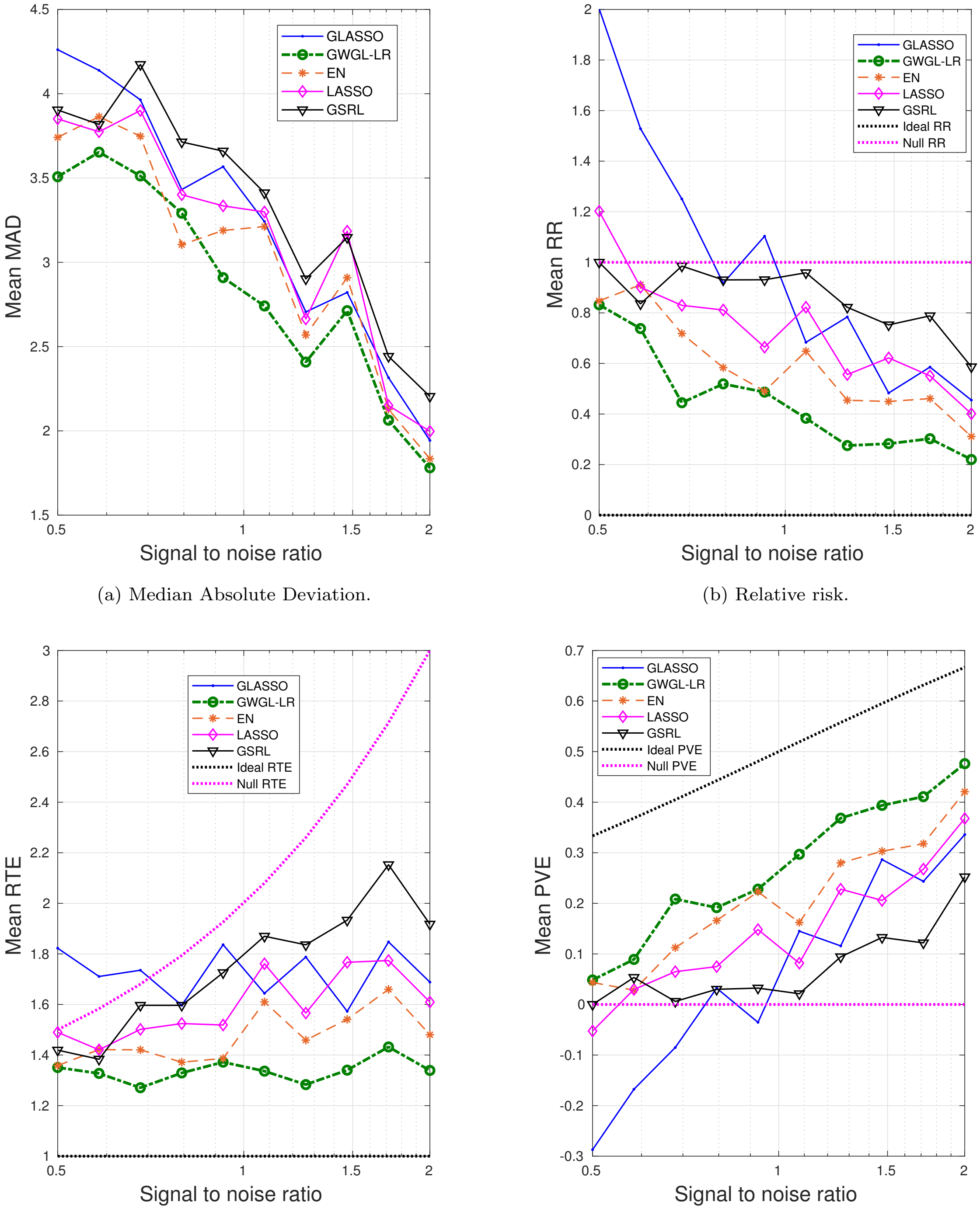
(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Fig. 1: The impact of within group correlation on the performance metrics, $q = 30\%$.

We plot two sets of graphs: (i) the performance metrics v.s. SNR, where SNR is equally spaced between 0.5 and 2 on a log scale, and ρ_w is set to 0.8 times a random noise uniformly distributed on the interval $[0.2, 0.4]$; and (ii) the performance metrics v.s. ρ_w, where ρ_w takes values in $(0.1, 0.2, \dots, 0.9)$, and SNR is fixed to 1. In the graphs for RR, RTE and PVE, we also plot the ideal scores, which are the values achieved by $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, and the null scores, which are the values achieved by $\hat{\boldsymbol{\beta}} = 0$. Results for $q = 30\%$ are in Figs. 2 and 1 while results for $q = 20\%$ can be found in Appendix B.



(a) Median Absolute Deviation.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Fig. 2: The impact of SNR on the performance metrics, $q = 30\%$.

To better highlight the benefits of GWGL-LR, we define the *Maximum Percentage Improvement (MPI)* to be the maximum percentage difference of the performance metrics between GWGL-LR and the best among all others. The MPI values for all metrics are shown in Tables 1 and 2 where we summarize the MPI brought about by our methods compared to other procedures, when varying the SNR and ρ_w, respectively. In all tables, the number outside the parentheses is the MPI value corresponding to each metric, while the number in the parentheses indicates the value of SNR/ρ_w at which the MPI is attained.

Table 1: MPI of all metrics when varying the SNR.

	MAD	RR	RTE	PVE
$q = 20\%$	13.7 (0.5)	41.4 (1.47)	13.1 (1.47)	68.9 (0.79)
$q = 30\%$	14.7 (1.08)	40.9 (1.08)	17 (1.08)	85.7 (0.68)

Table 2: MPI of all metrics when varying the within group correlation.

	MAD	RR	RTE	PVE
$q = 20\%$	8.2 (0.1)	80.5 (0.9)	31.8 (0.9)	145.4 (0.9)
$q = 30\%$	10.2 (0.1)	41.9 (0.1)	16.7 (0.1)	162.5 (0.1)

We summarize below our main findings from the results we have presented: (*i*) for all approaches, MAD and RR decrease as the data become less noisy. PVE increases when the noise is reduced; (*ii*) the GWGL-LR formulation has better prediction and estimation performance than all other approaches under consideration. When the within group correlation is varied, GWGL-LR shows a more stable performance; and (*iii*) the relative improvement of GWGL-LR over GLASSO is more significant for highly noisy data. Moreover, GWGL-LR generates more stable estimators than GLASSO.

4.2 Surgery Dataset

In this section we test our GWGL formulations on a real dataset obtained from the National Surgical Quality Improvement Program (NSQIP) containing medical records of patients who underwent a general surgical procedure. The dataset includes (*i*) baseline demographics; (*ii*) pre-existing comorbidity information; (*iii*) preoperative variables; (*iv*) index admission-related diagnosis and procedure information; (*v*) postoperative events and complications, and (*vi*) additional socioeconomic variables.

In our study, patients who underwent a general surgery procedure over 2011–2014 and were tracked by the NSQIP were identified. We will focus on two supervised learning models: (i) a linear regression model whose objective is to predict the post-operative hospital length of stay, and (ii) an LG model whose objective is to predict the re-hospitalization of patients within 30 days after discharge. Both models are extremely useful as they allow hospital staff to predict post-operative bed occupancy and prevent costly 30-day readmissions.

The post-processed datasets include a total of $2,275,452$ records, with 131 numerical predictors for the regression model and 132 for the classification model. The spectral clustering algorithm is used to group the predictors, with the number of groups specified as 67 based on a preliminary analysis.

For predicting the hospital length of stay, we report the mean (std.) of the out-of-sample MAD across 5 repetitions in Table 3. Our GWGL-LR formulation achieves the lowest mean MAD with a small variance; we improve the mean MAD by 7.30% over the best alternative. For longer hospital length of stay, this could imply 1 or 2 days improvement in prediction accuracy, which is both clinically and economically significant.

Table 3: The mean and standard deviation of MAD on the surgery data.

	GLASSO	GWGL-LR	EN	LASSO	GSRL
Mean (Std.)	0.17 (0.0007)	0.16 (0.001)	0.17 (0.0009)	0.17 (0.0009)	0.17 (0.0009)

For predicting the re-hospitalization of patients, we notice that the dataset is highly unbalanced, with only 6% of patients being re-hospitalized. To obtain a balanced training set, we randomly draw 20% patients from the positive class (re-hospitalized patients), and sample the same number of patients from the negative class, resulting in a training set of size $53,616$. All the remaining patients are assigned to the test dataset. All formulations achieve an average *out-of-sample ACC* (the prediction accuracy on the test dataset) around 0.62, an average *out-of-sample AUC* (Area Under the ROC Curve) of 0.83, and an average *logloss* on the test set ranging from 0.84 to 0.87. We define a new performance metric, called the *Within Group Difference (WGD)*, to measure the ability of the solution to induce group level sparsity.

$$\text{WGD}(\hat{\boldsymbol{\beta}}) \triangleq \frac{1}{|\{l : p_l \geq 2\}|} \sum_{l:p_l \geq 2} \frac{1}{\binom{p_l}{2}} \sum_{x_i, x_j \in \mathbf{x}^l} \left| \frac{\hat{\beta}_i - \hat{\beta}_j}{\mathbf{x}'_{,i}\mathbf{x}_{,j}} \right|,$$

where $|\{l : p_l \geq 2\}|$ denotes the cardinality of the set $\{l : p_l \geq 2\}$, and $\mathbf{x}'_{,i}\mathbf{x}_{,j}$ measures the sample correlation between predictors x_i and x_j. Theorem 3.1 implies that the higher the correlation, the smaller the difference between the coefficients, and thus, a smaller WGD value would suggest a stronger ability of grouped variable selection. Table 4 suggests that GWGL-LG encourages group level sparsity. From Table 5 (see the Appendix) we conclude that though LG-EN and LG-LASSO obtain the most parsimonious model at an individual level, GWGL-LG has a stronger ability to induce group level sparsity.

Table 4: The WGD of the estimators on the surgery data.

	LG	LG-LASSO	LG-Ridge	LG-EN	GWGL-LG
Mean (Std.)	23.93 (1.28)	16.28 (0.72)	23.38 (1.15)	16.26 (0.74)	5.04 (0.45)

5 Conclusions

We proposed a DRO formulation under the Wasserstein metric that recovers the GLASSO penalty for LAD and LG, through which we have established a connection between group-sparse regularization and robustness. We provided insights on the grouping effect of our estimators, which suggests the use of spectral clustering with the Gaussian similarity function to perform grouping on the predictors. We reported results from several experiments, showing that our formulations achieve more accurate and stable estimates, and have a stronger ability of inducing group level sparsity.

Acknowledgements We thank George Kasotakis, MD, MPH, for providing access to the surgery dataset. We also thank Taiyao Wang for help in processing this dataset. The research was partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under grants N00014-19-1-2571 and N00014-21-1-2844, by the NIH under grants R01 GM135930 and UL54 TR004130, by the DOE under grants DE-AR-0001282 and NETL-EE0009696, and by the Boston University Kilachand Fund for Integrated Life Science and Engineering.

Appendix

A Omitted Theoretical Results and Proofs

This section contains the theoretical statements and proofs that are omitted in Sections 2 and 3.

Proof of Theorem 2.1

Proof From the definition of the Wasserstein distance, $W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})$ is the optimal value of the following optimization problem:

$$\begin{aligned} \min_{\Pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \quad & \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ \Pi\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \\ \text{s.t.} \quad & \int_{\mathcal{Z}} \Pi\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z}, \\ & \int_{\mathcal{Z}} \Pi\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = q\mathbb{P}_{\text{out}}(\mathbf{z}_2) + (1-q)\mathbb{P}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z}. \end{aligned} \tag{16}$$

Similarly, $W_1(\mathbb{P}, \mathbb{P}_{\text{mix}})$ is the optimal value of the following optimization problem:

$$\begin{aligned} \min_{\Pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \quad & \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ \Pi\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \\ \text{s.t.} \quad & \int_{\mathcal{Z}} \Pi\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z}, \\ & \int_{\mathcal{Z}} \Pi\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = q\mathbb{P}_{\text{out}}(\mathbf{z}_2) + (1-q)\mathbb{P}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z}. \end{aligned} \tag{17}$$

We propose a decomposition strategy. For Problem (16), decompose the joint distribution Π as $\Pi = (1-q)S + qT$, where S and T are two joint distributions of $\mathbf{z}_1$ and $\mathbf{z}_2$. The first set of constraints in Problem (16) can be equivalently expressed as:

$$(1-q)\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 + q\int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = (1-q)\mathbb{P}_{\text{out}}(\mathbf{z}_1) + q\mathbb{P}_{\text{out}}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z},$$

which is satisfied if

$$\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \quad \int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z}.$$

The second set of constraints can be expressed as:

$$(1-q)\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 + q\int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = q\mathbb{P}_{\text{out}}(\mathbf{z}_2) + (1-q)\mathbb{P}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z},$$

which is satisfied if

$$\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = \mathbb{P}(\mathbf{z}_2), \quad \int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = \mathbb{P}_{\text{out}}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z}.$$

The objective function can be decomposed as:

$$\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ \Pi\big(d\mathbf{z}_1, d\mathbf{z}_2\big) = \ (1-q)\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ S\big(d\mathbf{z}_1, d\mathbf{z}_2\big) + q\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2) T\big(d\mathbf{z}_1, d\mathbf{z}_2\big).$$

Therefore, Problem (16) can be decomposed into the following two subproblems.

$$\text{Subproblem 1:} \qquad \begin{aligned} \min_{S \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \quad & \textstyle\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ S\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \\ \text{s.t.} \quad & \textstyle\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z}, \\ & \textstyle\int_{\mathcal{Z}} S\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = \mathbb{P}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z}. \end{aligned}$$

$$\text{Subproblem 2:} \qquad \begin{aligned} \min_{T \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \quad & \textstyle\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ T\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \\ \text{s.t.} \quad & \textstyle\int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \forall \mathbf{z}_1 \in \mathcal{Z}, \\ & \textstyle\int_{\mathcal{Z}} T\big(\mathbf{z}_1, \mathbf{z}_2\big) d\mathbf{z}_1 = \mathbb{P}_{\text{out}}(\mathbf{z}_2), \forall \mathbf{z}_2 \in \mathcal{Z}. \end{aligned}$$

Assume that the optimal solutions to the two subproblems are S^* and T^*, respectively, we know $\Pi_0 = (1-q)S^* + qT^*$ is a feasible solution to Problem (16). Therefore,

$$\begin{aligned} W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) &\leq \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1, \mathbf{z}_2)\ \Pi_0\big(d\mathbf{z}_1, d\mathbf{z}_2\big) \\ &= (1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P}) + qW_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{out}}) \\ &= (1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \end{aligned} \tag{18}$$

Similarly,

$$W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) \leq qW_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \tag{19}$$

(18) and (19) imply that

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) + W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) \leq W_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \tag{20}$$

On the other hand, based on the subadditivity of the Wasserstein metric, we have,

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) + W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) \geq W_1(\mathbb{P}_{\text{out}}, \mathbb{P}).$$

We thus conclude that

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) + W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) = W_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \tag{21}$$

To achieve the equality in (21), (18) and (19) must be equalities, i.e.,

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) = (1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P}),$$

and,

$$W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) = qW_1(\mathbb{P}_{\text{out}}, \mathbb{P}).$$

To see this, notice that if either (18) or (19) is a strict inequality, then (20) becomes a strict inequality, which contradicts (21). Thus,

$$\frac{W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})}{W_1(\mathbb{P}, \mathbb{P}_{\text{mix}})} = \frac{(1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P})}{qW_1(\mathbb{P}_{\text{out}}, \mathbb{P})} = \frac{1-q}{q}.$$

□

Proof of Theorem 2.2

Proof We will use Hölder's inequality, which we state for convenience.

Hölder's inequality: Suppose we have two scalars $p, q \geq 1$ and $1/p + 1/q = 1$. For any two vectors $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$,

$$\sum_{i=1}^n |a_i b_i| \leq \Big(\sum_{i=1}^n |a_i|^p\Big)^{1/p} \Big(\sum_{i=1}^n |b_i|^q\Big)^{1/q}.$$

The dual norm of $\|\cdot\|_{r,s}$ evaluated at some vector $\boldsymbol{\beta}$ is the optimal value of problem (22):

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}'\boldsymbol{\beta} \\ \text{s.t.} \quad & \|\mathbf{x}_{\mathbf{w}}\|_{r,s} \leq 1. \end{aligned} \tag{22}$$

We assume that $\boldsymbol{\beta}$ has the same group structure with $\mathbf{x}$, i.e., $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^L)$. Using Hölder's inequality, we can write

$$\mathbf{x}'\boldsymbol{\beta} = \sum_{l=1}^{L} (w_l \mathbf{x}^l)' \Big(\frac{1}{w_l}\boldsymbol{\beta}^l\Big) \leq \sum_{l=1}^{L} \|w_l \mathbf{x}^l\|_r \left\|\frac{1}{w_l}\boldsymbol{\beta}^l\right\|_q .$$

Define two new vectors in $\mathbb{R}^L$

$$\mathbf{x}_{new} = (\|w_1\mathbf{x}^1\|_r, \ldots, \|w_L\mathbf{x}^L\|_r), \quad \boldsymbol{\beta}_{new} = \left(\left\|\frac{1}{w_1}\boldsymbol{\beta}^1\right\|_q, \ldots, \left\|\frac{1}{w_L}\boldsymbol{\beta}^L\right\|_q\right).$$

Applying Hölder's inequality again to $\mathbf{x}_{new}$ and $\boldsymbol{\beta}_{new}$, we obtain:

$$\begin{aligned}
\mathbf{x}'\boldsymbol{\beta} &\leq \mathbf{x}'_{new}\boldsymbol{\beta}_{new} \\
&\leq \|\mathbf{x}_{new}\|_s \|\boldsymbol{\beta}_{new}\|_t \\
&= \Big(\sum_{l=1}^{L} \big(\|w_l\mathbf{x}^l\|_r\big)^s\Big)^{1/s} \left(\sum_{l=1}^{L} \left(\left\|\frac{1}{w_l}\boldsymbol{\beta}^l\right\|_q\right)^t\right)^{1/t} .
\end{aligned}$$

Therefore,

$$\mathbf{x}'\boldsymbol{\beta} \leq \|\mathbf{x}_{\mathbf{w}}\|_{r,s} \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{q,t} \leq \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{q,t},$$

due to the constraint $\|\mathbf{x}_{\mathbf{w}}\|_{r,s} \leq 1$. The result then follows. □

Proof of Theorem 2.3

Proof To derive a tractable reformulation of the DRO-LG problem (9), we borrow the idea from [8] and [14], which states that for any $\mathbb{Q} \in \Omega$,

$$\begin{aligned}
&\Big|\mathbb{E}^{\mathbb{Q}}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] - \mathbb{E}^{\hat{\mathbb{P}}_N}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big]\Big| \\
=&\Big|\int_{\mathcal{Z}} l_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1)\mathbb{Q}(d(\mathbf{x}_1, y_1)) - \int_{\mathcal{Z}} l_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\hat{\mathbb{P}}_N(d(\mathbf{x}_2, y_2))\Big| \\
=&\Big|\int_{\mathcal{Z}} l_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1)\int_{\mathcal{Z}} \Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)) - \int_{\mathcal{Z}} l_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\int_{\mathcal{Z}} \Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2))\Big| \\
\leq& \int_{\mathcal{Z}\times\mathcal{Z}} \big|l_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - l_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\big|\Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)),
\end{aligned} \tag{23}$$

where Π_0 is the optimal solution in the definition of the Wasserstein metric, i.e., it is the joint distribution of $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ that achieves the minimum mass transportation cost. Comparing (23) with the definition of the Wasserstein distance, we wish to bound the following *growth rate* of $l_{\boldsymbol{\beta}}(\mathbf{x}, y)$:

$$\frac{\big|l_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - l_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\big|}{s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))}, \ \forall (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),$$

in order to relate $\big|\mathbb{E}^{\mathbb{Q}}[l_{\boldsymbol{\beta}}(\mathbf{x}, y)] - \mathbb{E}^{\hat{\mathbb{P}}_N}[l_{\boldsymbol{\beta}}(\mathbf{x}, y)]\big|$ with $W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N)$. To this end, we define a continuous and differentiable univariate function $h(a) \triangleq \log(1 + \exp(-a))$, and apply the mean value theorem to it, which yields that for any $a, b \in \mathbb{R}$, $\exists c \in (a, b)$ such that:

$$\left|\frac{h(b) - h(a)}{b - a}\right| = \big|\triangledown h(c)\big| = \frac{e^{-c}}{1 + e^{-c}} \leq 1.$$

By noting that $l_{\boldsymbol{\beta}}(\mathbf{x}, y) = h(y\boldsymbol{\beta}'\mathbf{x})$, we immediately have:

$$\begin{aligned}
\big|l_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - l_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\big| &\le \big|y_1\boldsymbol{\beta}'\mathbf{x}_1 - y_2\boldsymbol{\beta}'\mathbf{x}_2\big| \\
&\le \|y_1\mathbf{x}_1 - y_2\mathbf{x}_2\|\|\boldsymbol{\beta}\|_* \\
&\le s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\|\boldsymbol{\beta}\|_*, \ \forall(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),
\end{aligned} \tag{24}$$

where the second step uses the Cauchy-Schwarz inequality, and the last step is due to the definition of the metric s in (8). Combining (24) with (23), it follows that for any $\mathbb{Q} \in \Omega$,

$$\begin{aligned}
\Big|\mathbb{E}^{\mathbb{Q}}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] - \mathbb{E}^{\hat{\mathbb{P}}_N}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big]\Big| &\le \|\boldsymbol{\beta}\|_* \int_{\mathcal{Z}\times\mathcal{Z}} s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)) \\
&= \|\boldsymbol{\beta}\|_* W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \le \epsilon\|\boldsymbol{\beta}\|_*.
\end{aligned}$$

Therefore, the DRO-LG problem can be reformulated as:

$$\inf_{\boldsymbol{\beta}} \mathbb{E}^{\hat{\mathbb{P}}_N}\big[l_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] + \epsilon\|\boldsymbol{\beta}\|_* = \inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \log\big(1 + \exp(-y_i\boldsymbol{\beta}'\mathbf{x}_i)\big) + \epsilon\|\boldsymbol{\beta}\|_*.$$

□

Prediction and Estimation Performance of the GWGL-LR Estimator

We are interested in two types of performance criteria: (1) *Prediction quality*, or out-of-sample performance, which measures the predictive power of the GWGL solutions on new, unseen samples. (2) *Estimation quality*, which measures the discrepancy between the GWGL solutions and the underlying unknown true coefficients.

We note that GWGL-LR is a special case of the Wasserstein DRO formulation derived in [8, Eq. 10], and thus the two types of performance guarantees derived in [8], one for generalization ability (prediction error), and the other for the discrepancy between the estimated and the true regression coefficients (estimation error), still apply to our GWGL-LR formulation.

We first establish a bound for the prediction bias of the solution to the GWGL-LR formulation, where the Wasserstein metric is induced by the weighted $(2, \infty)$-norm with weight $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M)$. The dual norm in this case is just the weighted $(2, 1)$-norm with weight $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L}, 1/M)$. Throughout this section we use $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$ to denote the true and estimated regression coefficient vectors, respectively. We first state several assumptions that are needed to establish the results.

Assumption A *The weighted* $(2, \infty)$*-norm of the uncertainty parameter* $(\mathbf{x}, y)$ *with weight* $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M)$ *is bounded above by* R *almost surely.*

Assumption B *For every feasible* $\boldsymbol{\beta}$*,* $\|(-\boldsymbol{\beta}^1, \ldots, -\boldsymbol{\beta}^L, 1)_{\mathbf{w}^{-1}}\|_{2,1} \le \bar{B}$*, where* $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L}, 1/M)$*.*

Let $\hat{\boldsymbol{\beta}}$ be an optimal solution to (7), obtained using the samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$. Suppose we draw a new i.i.d. sample $(\mathbf{x}, y)$. Using Theorem 3.3 in [8], Theorem A.1 establishes bounds on the error $|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|$.

Theorem A.1 *Under Assumptions A and B, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *with respect to the sampling,*

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

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

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

Theorem A.1 essentially says that with a high probability, the expected loss on new test samples using our GWGL-LR estimator can be upper bounded by the average loss in the training samples plus two terms that are related to the magnitude of the regularizer $\bar{B}$, the uncertainty level R, the confidence level δ, and converge to zero as $O(1/\sqrt{N})$. This result justifies the form of the regularizer used in (7) and guarantees a small generalization error of the GWGL-LR solution.

We next discuss the estimation performance of the GWGL-LR solution. Thm. A.2, a specialization of Thm. 3.11 in [8], provides a bound for the estimation bias in GWGL-LR. We first state the assumptions that are needed to establish the result.

Assumption C *The* ℓ_2 *norm of* $(-\boldsymbol{\beta}, 1)$ *is bounded above by* $\bar{B}_2$*.*

Assumption D *For some set*

$$\mathcal{A}(\boldsymbol{\beta}^*) := cone\{\mathbf{v}|\ \|(-\boldsymbol{\beta}^*, 1)_{\mathbf{w}^{-1}} + \mathbf{v}_{\mathbf{w}^{-1}}\|_{2,1} \leq \|(-\boldsymbol{\beta}^*, 1)_{\mathbf{w}^{-1}}\|_{2,1}\} \cap \mathbb{S}^{p+1}$$

and some positive scalar $\underline{\alpha}$*, the following holds,*

$$\inf_{\mathbf{v}\in\mathcal{A}(\boldsymbol{\beta}^*)} \mathbf{v}'\mathbf{Z}\mathbf{Z}'\mathbf{v} \geq \underline{\alpha},$$

where $\mathbf{Z} = [(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)]$ *is the matrix with columns* $(\mathbf{x}_i, y_i), i = 1, \ldots, N$*, and* $\mathbb{S}^{p+1}$ *is the unit sphere in the* $(p+1)$*-dimensional Euclidean space.*

Assumption E $(\mathbf{x}, y)$ *is a centered sub-Gaussian random vector, i.e., it has zero mean and satisfies the following condition:*

$$|||(\mathbf{x}, y)|||_{\psi_2} = \sup_{\mathbf{u}\in\mathbb{S}^{p+1}} \big|\big|\big|(\mathbf{x}, y)'\mathbf{u}\big|\big|\big|_{\psi_2} \leq \mu.$$

Assumption F *The covariance matrix of* $(\mathbf{x}, y)$ *has bounded positive eigenvalues. Set* $\boldsymbol{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$*; then,*

$$0 < \lambda_{min} \triangleq \lambda_{min}(\boldsymbol{\Gamma}) \leq \lambda_{max}(\boldsymbol{\Gamma}) \triangleq \lambda_{max} < \infty.$$

Definition 1 (Sub-Gaussian random variable) *A random variable* z *is sub-Gaussian if it is zero mean, and the* ψ_2*-norm defined below is finite, i.e.,*

$$|||z|||_{\psi_2} \triangleq \sup_{q\geq 1} \frac{(\mathbb{E}|z|^q)^{1/q}}{\sqrt{q}} < +\infty.$$

An equivalent property for sub-Gaussian random variables is that their tail distribution decays at least as fast as a Gaussian, i.e.,

$$\mathbb{P}(|z| \geq t) \leq 2\exp\{-t^2/C^2\}, \quad \forall t \geq 0,$$

for some constant C. A random vector $\mathbf{z} \in \mathbb{R}^{p+1}$ is sub-Gaussian if $\mathbf{z}'\mathbf{u}$ is sub-Gaussian for any $\mathbf{u} \in \mathbb{R}^{p+1}$. The ψ_2-norm of a vector $\mathbf{z}$ is defined as:

$$|||\mathbf{z}|||_{\psi_2} \triangleq \sup_{\mathbf{u}\in\mathbb{S}^{p+1}} \big|\big|\big|\mathbf{z}'\mathbf{u}\big|\big|\big|_{\psi_2},$$

where $\mathbb{S}^{p+1}$ denotes the unit sphere in the $(p+1)$-dimensional Euclidean space.

Definition 2 (Gaussian width) *For any set $\mathcal{A} \subseteq \mathbb{R}^{p+1}$, its Gaussian width is defined as:*

$$w(\mathcal{A}) \triangleq \mathbb{E}\Big[\sup_{\mathbf{u}\in\mathcal{A}} \mathbf{u}'\mathbf{g}\Big],$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a $(p+1)$-dimensional standard Gaussian random vector.

Theorem A.2 *Suppose the true regression coefficient vector is $\boldsymbol{\beta}^*$ and the solution to GWGL-LR is $\hat{\boldsymbol{\beta}}$. Under Assumptions A, C, D, E, and F, when the sample size $N \geq \bar{C}_1\bar{\mu}^4\mu_0^2(\lambda_{max}/\lambda_{min}) \cdot (w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3)^2$, with probability at least $1 - \exp(-C_2N/\bar{\mu}^4) - C_4\exp(-C_5^2(w(\mathcal{B}_u))^2/(4\rho^2))$,*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{min}} w(\mathcal{B}_u)\Psi(\boldsymbol{\beta}^*),$$

where $\bar{\mu} = \mu\sqrt{(1/\lambda_{min})}$; μ_0 is the ψ_2-norm of a standard Gaussian random vector $\mathbf{g} \in \mathbb{R}^{p+1}$; $w(\mathcal{A}(\boldsymbol{\beta}^*))$ is the Gaussian width (defined below) of $\mathcal{A}(\boldsymbol{\beta}^*)$ (cf. Assumption D); $w(\mathcal{B}_u)$ is the Gaussian width of $\mathcal{B}_u$, where $\mathcal{B}_u$ is the unit ball of the norm $\|\cdot\|_\infty$; $\rho = \sup_{\mathbf{v}\in\mathcal{B}_u}\|\mathbf{v}\|_2$; $\Psi(\boldsymbol{\beta}^*) = \sup_{\mathbf{v}\in\mathcal{A}(\boldsymbol{\beta}^*)}\|\mathbf{v}_{\mathbf{w}^{-1}}\|_{2,1}$; and $\bar{C}_1, C_2, C_4, C_5, \bar{C}$ are positive constants.

With Thm. A.2, we are able to provide bounds for performance metrics, such as the *Relative Risk (RR)*, *Relative Test Error (RTE)*, and *Proportion of Variance Explained (PVE)* [16]. All these metrics evaluate the accuracy of the regression coefficient estimates on a new test sample drawn from the same probability distribution as the training samples. Let $(\mathbf{x}_0, y_0)$ be such a test sample satisfying $y_0 = \mathbf{x}_0'\boldsymbol{\beta}^* + \eta_0$, where η_0 is random noise with zero mean and variance σ^2, and independent of the zero mean predictor $\mathbf{x}_0$. For a fixed set of training samples, let the solution to GWGL-LR be $\hat{\boldsymbol{\beta}}$. As in [16], define

$$\text{RR}(\hat{\boldsymbol{\beta}}) = \frac{\mathbb{E}(\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mathbf{x}_0'\boldsymbol{\beta}^*)^2}{\mathbb{E}(\mathbf{x}_0'\boldsymbol{\beta}^*)^2} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\mathbf{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{(\boldsymbol{\beta}^*)'\mathbf{\Sigma}\boldsymbol{\beta}^*},$$

where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{x}_0$, which is just the top left block of the matrix $\boldsymbol{\Gamma}$ in Assumption F. RTE is defined as:

$$\text{RTE}(\hat{\boldsymbol{\beta}}) = \frac{\mathbb{E}(y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}})^2}{\sigma^2} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{\sigma^2}.$$

PVE is defined as:

$$\text{PVE}(\hat{\boldsymbol{\beta}}) = 1 - \frac{\mathbb{E}(y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}})^2}{Var(y_0)} = 1 - \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{(\boldsymbol{\beta}^*)'\mathbf{\Sigma}\boldsymbol{\beta}^* + \sigma^2}.$$

Using Theorem A.2, we can bound the term $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ as follows:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \lambda_{max}(\boldsymbol{\Sigma})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq \lambda_{max}(\boldsymbol{\Sigma})\left(\frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{\text{min}}} w(\mathcal{B}_u)\Psi(\boldsymbol{\beta}^*)\right)^2, \tag{25}$$

where $\lambda_{max}(\boldsymbol{\Sigma})$ is the maximum eigenvalue of $\boldsymbol{\Sigma}$. Using (25), bounds for RR, RTE, and PVE can be readily obtained and are summarized in the following Corollary.

Corollary A.3 *Under the specifications in Theorem A.2, when the sample size*

$$N \geq \bar{C}_1\bar{\mu}^4\mu_0^2(\lambda_{max}/\lambda_{min})(w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3)^2,$$

with probability at least $1 - \exp(-C_2N/\bar{\mu}^4) - C_4\exp(-C_5^2(w(\mathcal{B}_u))^2/(4\rho^2))$,

$$RR(\hat{\boldsymbol{\beta}}) \leq \frac{\lambda_{max}(\boldsymbol{\Sigma})\left(\frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{min}} w(\mathcal{B}_u)\Psi(\boldsymbol{\beta}^*)\right)^2}{(\boldsymbol{\beta}^*)'\mathbf{\Sigma}\boldsymbol{\beta}^*},$$

$$RTE(\hat{\boldsymbol{\beta}}) \leq \frac{\lambda_{max}(\boldsymbol{\Sigma})\Big(\frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{min}}w(\mathcal{B}_u)\Psi(\boldsymbol{\beta}^*)\Big)^2 + \sigma^2}{\sigma^2},$$

$$PVE(\hat{\boldsymbol{\beta}}) \geq 1 - \frac{\lambda_{max}(\boldsymbol{\Sigma})\Big(\frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{min}}w(\mathcal{B}_u)\Psi(\boldsymbol{\beta}^*)\Big)^2 + \sigma^2}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^* + \sigma^2},$$

where all parameters are defined in the same way as in Theorem A.2.

Predictive Performance of the GWGL-LG Estimator

In this subsection we establish bounds on the prediction error of the GWGL-LG solution. Similar to [8], we will use the *Rademacher complexity* of the class of logloss (negative log-likelihood) functions to bound the generalization error. Two assumptions that impose conditions on the magnitude of the regularizer and the uncertainty level of the predictor are needed.

Assumption G *The weighted* $(2, \infty)$*-norm of* $\mathbf{x}$ *with weight* $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}})$ *is bounded above almost surely, i.e.,* $\|\mathbf{x}_{\mathbf{w}}\|_{2,\infty} \leq R_{\mathbf{x}}$.

Assumption H *The weighted* $(2, 1)$*-norm of* $\boldsymbol{\beta}$ *with* $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L})$ *is bounded above, namely,* $\sup_{\boldsymbol{\beta}} \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} = \bar{B}_1$.

Under these two assumptions, the logloss could be bounded via the Cauchy-Schwarz inequality.

Lemma A.4 *Under Assumptions G and H, it follows*

$$\log\big(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})\big) \leq \log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big), \quad \textit{almost surely.}$$

Now consider the following class of loss functions:

$$\mathcal{L} = \big\{(\mathbf{x}, y) \mapsto l_{\boldsymbol{\beta}}(\mathbf{x}, y) : l_{\boldsymbol{\beta}}(\mathbf{x}, y) = \log\big(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})\big),\ \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} \leq \bar{B}_1\big\}.$$

It follows from [8,4] that the empirical *Rademacher complexity* of $\mathcal{L}$, denoted by $\mathcal{R}_N(\mathcal{L})$, can be upper bounded by:

$$\mathcal{R}_N(\mathcal{L}) \leq \frac{2\log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big)}{\sqrt{N}}.$$

Then, applying Theorem 8 in [2], we have the following result on the prediction error of our GWGL-LG estimator.

Theorem A.5 *Let* $\hat{\boldsymbol{\beta}}$ *be an optimal solution to (11), obtained using* N *training samples* $(\mathbf{x}_i, y_i)$*,* $i = 1, \ldots, N$*. Suppose we draw a new i.i.d. sample* $(\mathbf{x}, y)$*. Under Assumptions G and H, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *with respect to the sampling,*

$$\mathbb{E}\big[\log\big(1 + \exp(-y\mathbf{x}'\hat{\boldsymbol{\beta}})\big)\big] \leq \frac{1}{N}\sum_{i=1}^N \log\big(1 + \exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})\big) + \frac{2\log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big)}{\sqrt{N}} + \log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big)\sqrt{\frac{8\log(2/\delta)}{N}}\ , \quad (26)$$

and for any $\zeta > \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}} + \log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big)\sqrt{\frac{8\log(2/\delta)}{N}}$,

$$\begin{aligned} &\mathbb{P}\Big(\log\big(1 + \exp(-y\mathbf{x}'\hat{\boldsymbol{\beta}})\big) \geq \frac{1}{N}\sum_{i=1}^N \log\big(1 + \exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})\big) + \zeta\Big) \\ &\leq \frac{\frac{1}{N}\sum_{i=1}^N \log\big(1 + \exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})\big) + \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}} + \log\big(1 + \exp(R_{\mathbf{x}}\bar{B}_1)\big)\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^N \log\big(1 + \exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})\big) + \zeta}\ . \end{aligned} \quad (27)$$

Theorem A.5 implies that the groupwise regularized LG formulation (11) yields a solution with a small generalization error on new i.i.d. samples.

Proof of Theorem 3.1 for GWGL-LR

Proof By the optimality condition associated with formulation (7), $\hat{\boldsymbol{\beta}}$ satisfies:

$$\mathbf{x}_{,i}'\text{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\sqrt{p_{l_1}}\frac{\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2}, \tag{28}$$

$$\mathbf{x}_{,j}'\text{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\sqrt{p_{l_2}}\frac{\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}, \tag{29}$$

where the sgn$(\cdot)$ function is applied to a vector elementwise. Subtracting (29) from (28), we obtain:

$$(\mathbf{x}_{,i} - \mathbf{x}_{,j})'\text{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\left(\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}\right).$$

Using the Cauchy-Schwarz inequality and $\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2 = 2(1-\rho)$, we obtain

$$\begin{aligned} D(i,j) &= \left|\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}\right| \\ &\leq \frac{1}{N\epsilon}\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2\|\text{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|_2 \\ &\leq \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon}. \end{aligned}$$

□

Proof of Theorem 3.1 for GWGL-LG

Proof By the optimality condition associated with formulation (11), $\hat{\boldsymbol{\beta}}$ satisfies:

$$\sum_{k=1}^{N}\frac{\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1+\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}y_k x_{k,i} = N\epsilon\sqrt{p_{l_1}}\frac{\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2}, \tag{30}$$

$$\sum_{k=1}^{N}\frac{\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1+\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}y_k x_{k,j} = N\epsilon\sqrt{p_{l_2}}\frac{\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}, \tag{31}$$

where $x_{k,i}$ and $x_{k,j}$ denote the i-th and j-th elements of $\mathbf{x}_k$, respectively. Subtracting (31) from (30), we get:

$$\sum_{k=1}^{N}\frac{\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1+\exp(-y_k\mathbf{x}_k'\hat{\boldsymbol{\beta}})}\big(y_k x_{k,i} - y_k x_{k,j}\big) = N\epsilon\left(\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}\right). \tag{32}$$

Note that the LHS of 32 can be written as $\mathbf{v}_1'\mathbf{v}_2$, where

$$\mathbf{v}_1 = \left(\frac{\exp(-y_1\mathbf{x}_1'\hat{\boldsymbol{\beta}})}{1+\exp(-y_1\mathbf{x}_1'\hat{\boldsymbol{\beta}})}, \ldots, \frac{\exp(-y_N\mathbf{x}_N'\hat{\boldsymbol{\beta}})}{1+\exp(-y_N\mathbf{x}_N'\hat{\boldsymbol{\beta}})}\right),$$

$$\mathbf{v}_2 = \big(y_1(x_{1,i} - x_{1,j}), \ldots, y_N(x_{N,i} - x_{N,j})\big).$$

Using the Cauchy-Schwarz inequality and $\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2 = 2(1-\rho)$, we obtain

$$\begin{aligned} D(i,j) &= \left|\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}\right| \\ &\leq \frac{1}{N\epsilon}\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2 \\ &\leq \frac{1}{N\epsilon}\sqrt{N}\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2 = \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon}. \end{aligned}$$

□

B Omitted Numerical Results

This section contains the experimental setup and results that are omitted in Section 4.

Omitted Results in Section 4.1

Hyperparameter Tuning

All the penalty parameters are tuned using a separate validation dataset. Specifically, we divide all the N training samples into two sets, dataset 1 and dataset 2 (validation set). For a pre-specified range of values for the penalty parameters, dataset 1 is used to train the models and derive $\hat{\boldsymbol{\beta}}$, and the performance of $\hat{\boldsymbol{\beta}}$ is evaluated on dataset 2. We choose the penalty parameter that yields the minimum unpenalized loss of the respective approaches on the validation set. As to the range of values for the tuned parameters, we borrow ideas from [16], where the LASSO was tuned over 50 values ranging from $\lambda_m \triangleq \|\mathbf{X}'\mathbf{y}\|_\infty$ to a small fraction of λ_m on a log scale. In our experiments, this range is properly adjusted for the GLASSO estimators. Specifically, for GWGL and GSRL, the tuning range is: $\sqrt{\exp(\text{lin}(\log(0.005 \cdot \|\mathbf{X}'\mathbf{y}\|_\infty), \log(\|\mathbf{X}'\mathbf{y}\|_\infty), 50))/\max(p_1, \ldots, p_L)}$, where the function $\text{lin}(a, b, n)$ takes in scalars a, b and n (integer) and outputs a set of n values equally spaced between a and b; the exp function is applied elementwise to a vector. Compared to LASSO, the values are scaled by $\max(p_1, \ldots, p_L)$, and the square root operation is due to the ℓ_1-loss function, or the square root of the ℓ_2-loss used in these formulations. For the GLASSO with ℓ_2-loss, the range is: $\exp(\text{lin}(\log(0.005 \cdot \|\mathbf{X}'\mathbf{y}\|_\infty), \log(\|\mathbf{X}'\mathbf{y}\|_\infty), 50))/\sqrt{\max(p_1, \ldots, p_L)}$.

Implementation of Spectral Clustering

In our implementation, the k-nearest neighbor similarity graph is constructed, where we connect $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ with an undirected edge if $\mathbf{x}_{,i}$ is among the k-nearest neighbors of $\mathbf{x}_{,j}$ (in the sense of Euclidean distance) *or* if $\mathbf{x}_{,j}$ is among the k-nearest neighbors of $\mathbf{x}_{,i}$. The parameter k is chosen such that the resulting graph is connected. Recall that we use the Gaussian similarity function

$$\text{Gs}(\mathbf{x}_{,i}, \mathbf{x}_{,j}) \triangleq \exp\big(-\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2/(2\sigma_s^2)\big), \tag{33}$$

to construct the graph. The scale parameter σ_s in (33) is set to the mean distance of a point to its k-th nearest neighbor [30]. We assume that the number of clusters is known in order to perform spectral clustering, but in case it is unknown, the eigengap heuristic [30] can be used, where the goal is to choose the number of clusters c such that all eigenvalues $\lambda_1, \ldots, \lambda_c$ of the graph Laplacian are very small, but λ_{c+1} is relatively large.

The number of dropped groups/features on the surgery data

These results are in Table 5.

The Impact on the Performance Metrics when $q = 20\%$

See Figs. 3 and 4.

Table 5: The number of dropped groups/features on the surgery data.

	LG	LG-LASSO	LG-Ridge	LG-EN	GWGL-LG
No. of dropped groups	1	6	2	10	16
No. of dropped features	2	24	2	25	19

Omitted Results in Section 4.2

Pre-processing the Dataset

Data were pre-processed as follows: (*i*) categorical variables (such as race, discharge destination, insurance type) were numerically encoded and units homogenized; (*ii*) missing values were replaced by the mode; (*iii*) all variables were normalized by subtracting the mean and divided by the standard deviation; (*iv*) patients who died within 30 days of discharge or had a postoperative length of stay greater than 30 days were excluded.

References

1. Sergey Bakin. Adaptive regression and model selection in data mining problems. 1999.
2. Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
3. Dimitris Bertsimas and Martin S Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 2017.
4. Dimitris Bertsimas, Vishal Gupta, and Ioannis Ch Paschalidis. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153(2):595–633, 2015.
5. Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. *arXiv preprint arXiv:1705.04241*, 2017.
6. Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013.
7. Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root LASSO: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.
8. Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.
9. Ruidi Chen and Ioannis Ch Paschalidis. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.
10. Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

11. John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

12. Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Available at Optimization Online*, 2015.

13. Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

14. Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

15. Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

16. Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

17. Junzhou Huang, Tong Zhang, et al. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

18. Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group LASSO with overlap and graph LASSO. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

19. Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011.

20. Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

21. Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

22. Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

23. Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group LASSO with overlaps: the latent group LASSO approach. *arXiv preprint arXiv:1110.0413*, 2011.

24. Volker Roth and Bernd Fischer. The group-LASSO for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM, 2008.

25. Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

26. Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *arXiv preprint arXiv:1710.10016*, 2017.

27. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

28. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group LASSO. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

29. Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

30. Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

31. Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and LASSO. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

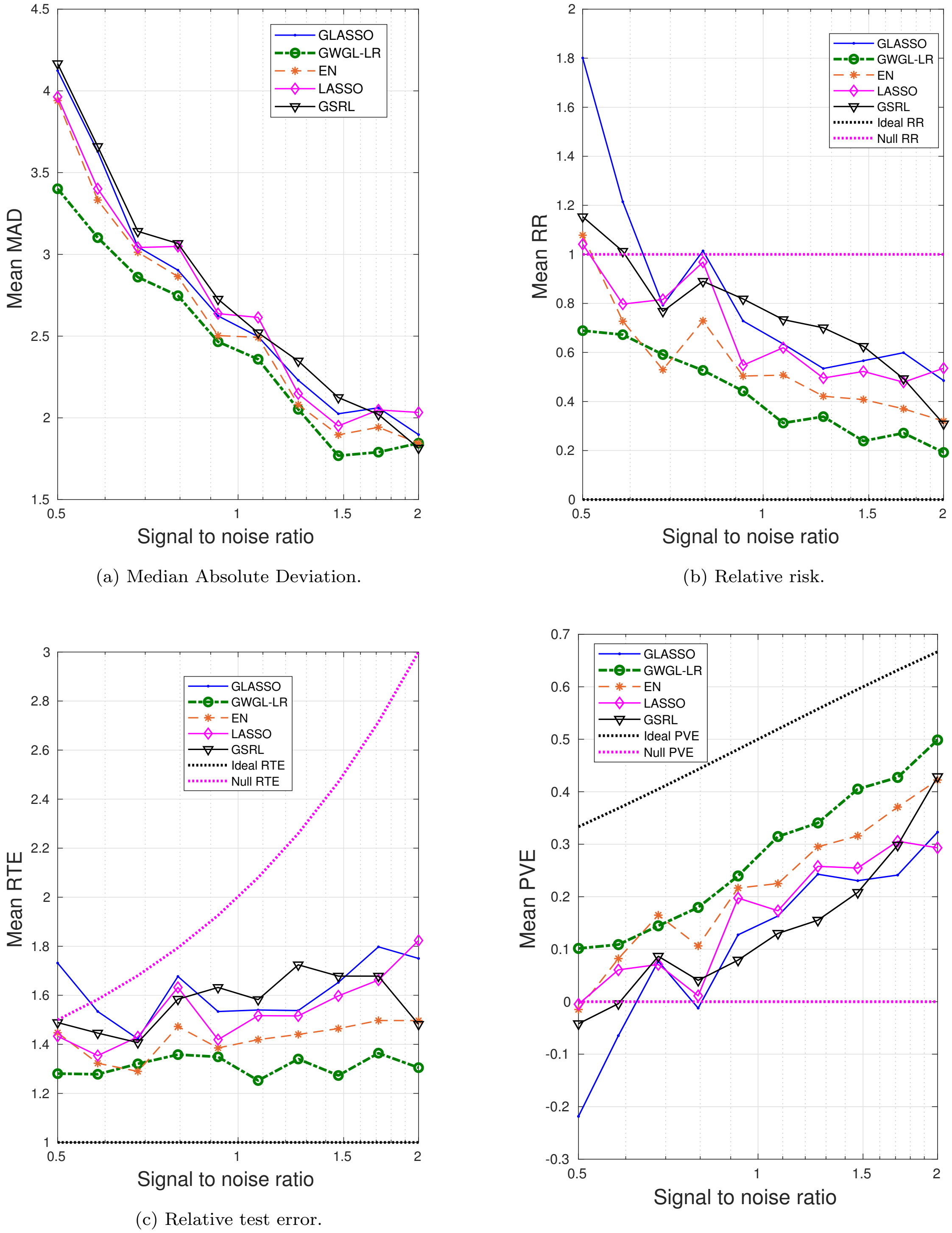
32. Wenzhuo Yang and Huan Xu. A unified robust regression model for LASSO-like algorithms. In *International Conference on Machine Learning*, pages 585–593, 2013.

33. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

34. Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

35. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

36. S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.



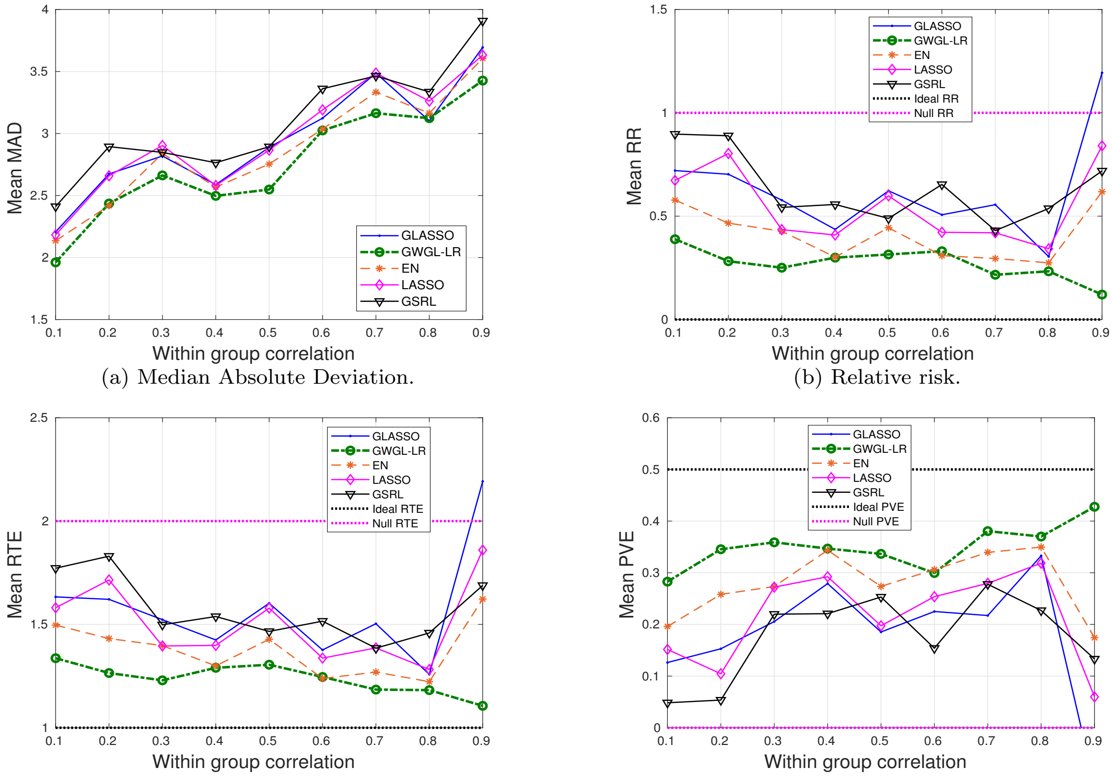
(a) Median Absolute Deviation.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Fig. 3: The impact of SNR on the performance metrics, $q = 20\%$.



(a) Median Absolute Deviation.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Fig. 4: The impact of within group correlation on the performance metrics, $q = 20\%$.