Journal of Machine Learning Research 23 (2022) 1-50 Submitted 7/21; Revised 2/22; Published 4/22

Meta-analysis of heterogeneous data: integrative sparse regression in high-dimensions

Subha Maity SMAITY@UMICH.EDU
Yuekai Sun YUEKAI@UMICH.EDU
Moulinath Banerjee MOULIB@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI

Editor: Qiang Liu

Abstract

We consider the task of meta-analysis in high-dimensional settings in which the data sources are similar but non-identical. To borrow strength across such heterogeneous datasets, we introduce a global parameter that emphasizes interpretability and statistical efficiency in the presence of heterogeneity. We also propose a one-shot estimator of the global parameter that preserves the anonymity of the data sources and converges at a rate that depends on the size of the combined dataset. For high-dimensional linear model settings, we demonstrate the superiority of our identification restrictions in adapting to a previously seen data distribution as well as predicting for a new/unseen data distribution. Finally, we demonstrate the benefits of our approach on a large-scale drug treatment dataset involving several different cancer cell-lines[1].

Keywords: Robust statistics, sparse-regression, meta-analysis

1. Introduction

We consider the task of synthesizing information from multiple datasets. This is usually done to enhance statistical power by increasing the effective sample size. However, as seen in many pooled studies, the inherent heterogeneity among the studies must be managed carefully. For example, in pooled genome-wide association studies (GWAS), differences in study populations and measurement methods may confound the association between the biomarkers and the outcome (Leek et al., 2010). To obtain meaningful conclusions, practitioners must properly model the heterogeneity in pooled studies.

In this paper, we consider the task of fitting high-dimensional regression models to m similar, but non-identical datasets. To borrow strength across the datasets, we must distinguish between the common and idiosyncratic parts of the datasets. Let $\theta_k^* \in \mathbf{R}^d$ be the vector of regression coefficients for the k-th dataset. Without loss of generality, we posit that the θ_k^*'s are related through an additive structure:

$$\theta_k^* = \theta_0^* + \delta_k^*, \quad k \in [m]. \tag{1.1}$$

1. Codes are available in https://github.com/smaityumich/MrLasso.

This is a decomposition of the θ_k^*'s into common and idiosyncratic parts: the *global parameter* $\theta_0^* \in \mathbf{R}^d$ represents the common part in the datasets, while the δ_k^*'s represent the idiosyncratic parts. By itself, this decomposition does not identify the common and idiosyncratic parts because the decomposition is not unique: for any $\delta \in \mathbf{R}^d$, $(\theta_0^*, \delta_1^*, \ldots, \delta_m^*)$ and $(\theta_0^* - \delta, \delta_1^* + \delta, \ldots, \delta_m^* + \delta)$ are different decompositions of the same θ_k^*'s. Thus it is necessary to impose additional identification restrictions on θ_0^* and the δ_k^*'s. Note that this identification issue is not specific to the additive decomposition (1.1); any parameterization of the θ_k^*'s that introduces a global parameter (to model the common parts of the datasets) is overparameterized/underdetermined.

As we shall see, the choice of identification restriction is crucial to the efficacy of meta-analysis; an improperly defined global parameters may *totally negate its benefits*. For example, consider the common ANOVA identification restriction:

$$\textstyle\sum_{k=1}^m \delta_k^* = 0.$$

This restriction implies that the global parameter is the average of the θ_k^*'s:

$$\theta_0^* \triangleq \tfrac{1}{m} \textstyle\sum_{k=1}^m \theta_k^*. \tag{1.2}$$

In some cases (*e.g.* when the θ_k^*'s have disjoint supports), the global parameter can be up to m-times denser than the local parameters, which negates any statistical efficiency gains from borrowing strength across the datasets. We defer a more comprehensive discussion of the inadequacies of common identification restrictions to Section 2.

In this paper, we define the global parameter as

$$(\theta_0^*)_j \triangleq \arg\min_{\mu_j \in \mathbf{R}} \textstyle\sum_{k=1}^m \Psi((\theta_k^*)_j - \mu_j), \quad j \in [d], \tag{1.3}$$

where Ψ is a *re-descending loss function* (Huber, 1964). Intuitively, this definition separates $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$ into outlier and inliers and defines $(\theta_0^*)_j$ as the average of the inliers. Thus the definition identifies the location of the inliers as the common part of the datasets. This identification restriction is motivated by an ϵ-contamination model for the θ_k^*'s:

$$(\theta_k^*)_j \sim (1-\epsilon)F_j + \epsilon G_j, \tag{1.4}$$

where $\epsilon < \frac{1}{2}$ is the fraction of outliers, F_j is the distribution of the inliers, and G_j is the distribution of the outliers (among the j-th regression coefficients). The two main benefits of this identification restriction are

1. *interpretability:* if a majority of $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$ are zero, then $(\theta_0^*)_j$ is zero. More generally, if there is a majority value among $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$ (i.e. a common value shared by more than half of $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$), then $(\theta_0^*)_j$ is the majority value. Furthermore, if most of the local parameter coordinates $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$) are in a *bulk* (considered as inliers) and a few of them are significantly different (outliers), then $(\theta_0^*)_j$ is the average of only the inliers. This ensures that the global parameter is *sparse* and *interpretable* as the common part of the local datasets.
2. *statistical efficiency:* θ_0^* can be estimated at a faster rate than the θ_k^*'s. This ensures meta-analysis leads to gains in statistical efficiency.

We note that the ANOVA identification restriction does not share these two benefits.

The rest of this paper is organized as follows. In Section 2, we describe the pitfalls of several common definitions of the global parameter in integrative studies and propose a robustness-based definition to avoid such pitfalls. We design an estimator of the robustness-based global parameter in Section 3 and prove in Section 4 that it converges at a rate *that depends on the total number of samples in all the datasets*. Finally, we demonstrate the benefits of our approach in predicting the response of rare cancers to therapeutics with the Cancer Cell Line Encyclopedia (CCLE).

1.1 Related work

The goal of meta-analysis is borrowing strength from different datasets, and the most common approach is a two-step method in which the local parameters are first estimated from their respective datasets and then combined with (say) a fixed-effects model (Hedges and Olkin, 2014) or a random-effects model (DerSimonian and Laird, 1986). In the high-dimensional setting, we must also perform variable selection to reduce the prediction/estimation error and ensure that the fitted model is interpretable.

Gross and Tibshirani (2016); Asiaee et al. (2018) studied the case of heterogeneous linear regression models, where it is assumed that the underlying distribution of data sets are not the same. The former focuses on the prediction aspects of the problem, while the latter deals with the estimation aspects. Both papers reduce the problem of meta-analysis into a single lasso exercise, while the latter uses a version of the gradient descent algorithm to estimate parameters. Heterogeneity in the Cox model was studied by Cheng et al. (2015) where the likelihood was maximized using a suitable lasso problem. However, all these methods required the full data sets for analysis to be available on a single platform. This raises the question of communication efficiency and privacy concerns pertaining to the data sets. Cai et al. (2021) proposed a communication-efficient integrative analysis for high-dimensional heterogeneous data which addresses the issue of privacy preservation. In their two-step estimation procedure, they used lasso estimates and covariance matrices to obtain an estimator for the shared parameter, which, in a nutshell, is the average of the debiased lasso estimates.

There is a line of work on distributed statistical estimation and inference, *e.g.*, Lee et al. (2017); Battey et al. (2018); Jordan et al. (2016), which is distinguished from our work by the additional assumption of no heterogeneity: *i.e.* the datasets are identically distributed. In this line of work, there are two general approaches: averaging local estimates of the parameters (Lee et al., 2017; Battey et al., 2018) and averaging local estimates of the score/sufficient statistic (Jordan et al., 2016). Although averaging the score has computational benefits over averaging the parameters, the latter is more amenable to meta-analysis because modeling the heterogeneity in the parameters is easier than modeling heterogeneity in the score.

2. Identifying the global parameter

We now quickly review the developments thus far. To distinguish between the common and idiosyncratic parts of the local parameters, we model the *local parameters* $\theta_1^*, \ldots, \theta_m^*$ with an additive model:

$$\theta_k^* = \theta_0^* + \delta_k^*, \quad k \in [m].$$

By itself, this additive model does not *identify* the global parameter θ_0^* because the model is overparameterized. Thus it is necessary to impose additional identification restrictions to uniquely identify the global parameter. As we saw in Section 1, the choice of identification restriction is crucial to the statistical efficiency of meta-analysis. Recall we look for two properties in an identification restriction: interpretability and statistical efficiency. We describe them here again for the reader's convenience:

1. *interpretability:* if there is a majority value among $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$ (i.e. a common value shared by more than half of $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$), then $(\theta_0^*)_j$ is the majority value. This encourages sparsity in the global parameter θ_0^* because if a majority of $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$ are zero, then $(\theta_0^*)_j$ is zero.
2. *statistical efficiency:* θ_0^* can be estimated at a faster rate than the θ_k^*'s. This ensures meta-analysis leads to gains in statistical efficiency.

In the rest of this section, we show that standard identification restrictions do not satisfy the two preceding properties. After describing the inadequacies of standard identification restrictions, we present two that satisfy our desiderata.

2.1 Inadequacies of standard identification restrictions

Mean: The usual approach to borrowing strength across heterogeneous datasets is to model the variation among the local parameters with a prior and consider the (hyper)parameter of the prior as the global parameter. The celebrated empirical Bayes approach is a prominent example. One of the simplest and most widely used priors is the Gaussian prior:

$$(\theta_k^*)_j \sim N((\theta_0^*)_j, \sigma_0^2).$$

The standard estimator of $(\theta_0^*)_j$ is $\frac{1}{m}\sum_{k=1}^m (\theta_k^*)_j$, which leads to the mean identification restriction (1.2). Unfortunately, this identification restriction does not satisfy our desiderata. Intuitively, the issue is it aggregates all local parameters in the definition of the global parameter regardless of whether a local parameter is similar to the other local parameters. This causes $(\theta_0^*)_j$ to lack interpretability because even a single non-zero local parameter is enough to nudge the global parameter away from zero. In the worst case (when the sparsity patterns of the local parameters are disjoint), the sparsity of the global parameter can be m times the sparsity of the local parameters. This extra complexity of the global parameter negates any gains in statistical efficiency from meta analysis.

Median: To address the sensitivity of the mean identification restriction to outliers, we consider identifying the global parameter with more robust centrality measures. One obvious choice is the median:

$$\theta_0^* \triangleq \arg\min_{\theta_0 \in \mathbf{R}^d} \sum_{k=1}^m \|\theta_k^* - \theta_0\|_1. \tag{2.1}$$

Unfortunately, in certain scenarios, the median may *fail to borrow strength* across datasets, making it statistically inefficient. If the local parameters are well-separated (see Figure 1), then the global parameter is one of the local parameters. In this case, it is impossible to estimate the global parameter at a faster rate than the equivalent local parameter. We provide a detailed example in Appendix A (see Example 1).

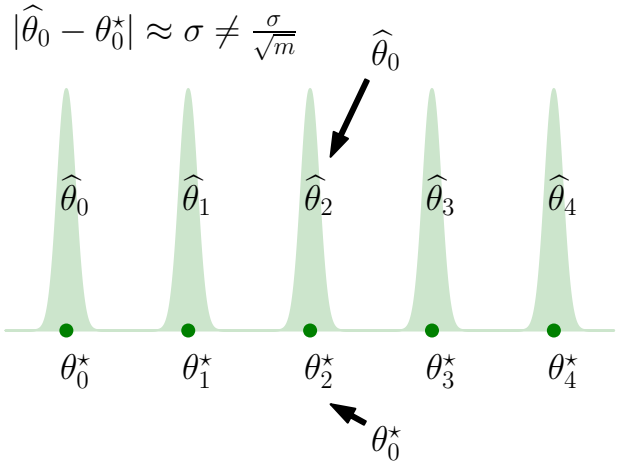

Figure 1: Median fails to borrow strength form well-separated local parameter estimates.

We note that Asiaee et al. (2018) and Gross and Tibshirani (2016) both considered identification restrictions similar to (2.1). In particular, they studied the estimator

$$(\widehat{\theta}_0, \widehat{\delta}_1, \ldots, \widehat{\delta}_m) \in \arg\min_{(\theta_0, \delta_1, \ldots, \delta_m)} \left\{ \begin{array}{l} \frac{1}{2N} \sum_{k=1}^m \|\mathbf{Y}_k - \mathbf{X}_k(\theta_0 + \delta_k)\|_2^2 \\ \quad + \frac{1}{m}(\lambda_0 \|\theta_0\|_1 + \sum_{k=1}^m \lambda_k \|\delta_k\|_1) \end{array} \right\},$$

where $\mathbf{X}_k \in \mathbf{R}^{n_k \times d}$ and $\mathbf{Y}_k \in \mathbf{R}^{n_k}$ are the features and responses in the k-th dataset and $N = \sum_{k=1}^m n_k$ is the total sample size. This estimator implicitly enforces the identification restriction

$$\theta_0^* \triangleq \arg\min_{\theta_0} \left(\sum_{k=1}^m \|\theta_k^* - \theta_0\|_1 + c\|\theta_0\|_1\right)$$

for some $c > 0$, which, unfortunately, suffers from a similar drawback as (2.1). We refer to Result 11 in Appendix A for the details. We note that this issue is reflected in Asiaee et al. (2018)'s theoretical results. In particular, they showed that

$$\|\widehat{\theta}_0 - \theta_0^*\|_2 + \sum_{k=1}^m \frac{1}{\sqrt{m}} \|\widehat{\delta}_k - \delta_k^*\|_2 \lesssim_P \max_k m \left(\frac{\max_k s_k \log d}{N}\right)^{\frac{1}{2}},$$

where s_k is the sparsity of θ_k^*. If we assume $\theta_0^*, \delta_1^*, \ldots, \delta_m^*$ are all s-sparse, then the right side simplifies to $\sqrt{m}(\frac{s \log d}{n})^{\frac{1}{2}}$. While this is the fastest rate of convergence we expect for the left side, it suggests that the estimator of the global parameter $\widehat{\theta}_0$ converges at a rate that depends on the local sample size.

Huber loss: As an alternative to the median, we consider minimizing Huber's loss (Huber, 1964). Huber's loss function is defined as

$$L_\eta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \eta \\ \eta\left(|x| - \frac{1}{2}\eta\right) & \text{otherwise,} \end{cases}$$

for some $\eta > 0$, and the global parameter is defined (coordinate-wise) as:

$$(\theta_0^*)_j \triangleq \arg\min_{\theta_0} \sum_{k=1}^m L_\eta((\theta_k^*)_j - \theta_0) \tag{2.2}$$

We hope that the quadratic part of Huber's loss allows it to borrow strength effectively across datasets. Unfortunately, this leads to a loss of interpretability. Consider a problem in which all but one local parameter values are identically zero. It is possible to show that the global parameter is non-zero for any $\eta > 0$ (see Example 2 in Appendix A).

2.2 Two interpretable and statistically efficient global parameters

The three examples presented in this subsection are hardly exhaustive. We list them here to illustrate the delicacy of the choice of identification restriction. In the rest of this section, we provide two examples of identification restrictions that satisfy our desiderata.

Both the examples have hyperparameters, that one must choose carefully, depending on the properties of the local datasets to satisfy our desiderata. At first, this seems unusual because the global parameter depends on the hyperparameters. However, the global parameter is defined by the data scientist to facilitate meta-analysis, and the hyperparameters *reflect the discretion of the data scientist*. Thus the dependence of the global parameter on the hyperparameters is natural. From another perspective, different values of hyperparameter lead to possible global parameters. However, not all the global parameters are interpretable and neither can they be estimated in a statistically efficient manner. We only consider global parameters that have these desirable properties.

Re-descending loss: In this paper, we define the global parameter as the minimum of a re-descending loss function (Huber, 1964). More concretely, we consider the quadratic re-descending loss function: $\Psi_\eta(x) \triangleq \min\{x^2, \eta^2\}$, and we define the global parameter as

$$(\theta_0^*)_j \triangleq \arg\min_{x \in \mathbf{R}} \sum_{k=1}^m \Psi_{\eta_j}((\theta_k^*)_j - x)\,. \tag{2.3}$$

It is possible to derive similar results for other re-descending loss functions, but we focus on the quadratic re-descending loss function here. As we shall see, $(\theta_0^*)_j$ is not only interpretable but also statistically efficient.

First, we check that (2.3) leads to an interpretable global parameter. The quadratic re-descending loss function has the property that its derivative vanishes outside $[-\eta_j, \eta_j]$, so (2.3) ignores any local parameters that are outside this interval. Thus local parameters that are far from the bulk of the local parameter values are considered outliers and ignored by (2.3), thereby ensuring that the global parameter is interpretable. For example, if there is a majority value among $(\theta_1^*)_j, \ldots, (\theta_m^*)_j$, then for a suitable choice of η_j, (2.3) ignores all local parameters that are different from the majority value.

Second, we argue that it is possible to estimate $(\theta_0^*)_j$ at a fast rate. Consider $(\theta_0^*)_j$ as an M-estimator in which the effective sample size is the number of local parameters in bulk (not considered outliers). This suggests that it is possible to estimate $(\theta_0^*)_j$ at a rate that depends on the number of local parameters in the bulk. As we shall see, the simple approach of replacing the local parameters with their estimates in (2.3) leads to an estimator that achieves this goal. We refer to Lemma 6 for a formal statement of a result to this effect.

We note that the choice of η_j plays a crucial role in the interpretability and statistical efficiency of the global parameter $(\theta_0^*)_j$. If η_j is very large, then none of the local parameter

values will be identified as outliers and (2.3) is equivalent to the mean identification restriction. This leads to a global parameter that is sensitive to outlier local parameters. On the other hand, if η_j is small, then (2.3) ignores most local parameter values because it considers them as outliers. This reduces the statistical efficiency of borrowing strength across datasets because most datasets are ignored. Thus η_j must be chosen in a way that balances interpretability and statistical efficiency.

We note that the quadratic re-descending loss has a close connection with the ϵ-contaminated model (1.4), where the inliers parameters are normally distributed:

$$(\theta_k^*)_j \sim (1-\epsilon)\mathcal{N}\big((\theta_0^*)_j, \sigma^2\big) + \epsilon G_j\,. \tag{2.4}$$

In the absence of contamination ($\epsilon = 0$), it is not hard to check that minimizing the squared loss function leads to the maximum likelihood estimator for the global parameter θ_0^*. In the presence of outliers, we can avoid the corrupting effects of the outliers by using a robust version of the squared loss function. There are many choices of such robust loss; one such choice is a re-descending loss function (Hampel, 2005).

Quadratic + ℓ_1 loss As an alternative to the re-descending loss, we could also consider minimizing a convex combination of quadratic and ℓ_1 loss functions:

$$\theta_0^* \triangleq \arg\min_{\theta\in\mathbf{R}} \sum_{k=1}^m \tfrac{1}{1+\lambda}\big\{\lambda\|\theta_k^* - \theta\|_1 + \tfrac{1}{2}\|\theta_k^* - \theta\|_2^2\big\}\,. \tag{2.5}$$

This identification restriction combines the mean and the median identification restrictions, but in a different way than Huber's loss function. It is not hard to check that this identification restriction satisfies our desiderata. The quadratic part of (2.5) allows us to borrow strength across datasets, while it is possible to pick λ in a way such that θ_0^* is interpretable. For example, if all but one of the local parameters are zero, it is possible to pick λ large enough so that the global parameter is zero, in contrast to the Huber loss function. That said, we focus on the re-descending loss here because it has better empirical performance (see Section 5).

3. Communication-efficient data enriched regression

In this section, we suggest a privacy-preserving communication-efficient estimator for the global parameter θ_0^* which borrows strength over different datasets. The privacy concern limits us to communicate with the datasets only through some summary statistics. So, the high-level idea to estimate θ_0^* is to start with some estimator of the local parameters computed from the datasets. Then the global parameter is estimated only using local estimates, without any further communication among datasets.

We describe the debiased lasso estimator for local parameters in the set-up of ℓ_1 regularized M-estimators. The case of the linear regression model can be considered as a special case. Let $\rho(y,a)$ be a loss function, which is convex in a, and $\dot\rho$, $\ddot\rho$ be its derivatives with respect to a. That is

$$\dot\rho(y,a) = \frac{d}{da}\rho(y,a), \quad \ddot\rho(y,a) = \frac{d^2}{da^2}\rho(y,a).$$

Define $\ell_k(\theta_k) = \frac{1}{n_k}\sum_{i=1}^{n_k}\rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta_k)$, where the sum is only over the pairs on dataset k. The lasso estimator for local parameter θ_k^* is given by

$$\widetilde{\theta}_k := \arg\min_{\theta\in\mathbf{R}^d}\ell_k(\theta) + \lambda_k\|\theta\|_1. \tag{3.1}$$

Since averaging only reduces variance, not bias, we gain (almost) nothing by averaging the biased lasso estimators. That is, the MSE of the naive average estimator is of the same order as that of the local estimators. The key to overcoming the bias of the averaged lasso estimator is to "debias" the lasso estimators before averaging.

The *debiased lasso estimator* as in van de Geer et al. (2014) is

$$\widetilde{\theta}_k^d = \widetilde{\theta}_k - \widehat{\Theta}_k\nabla\ell_k(\widetilde{\theta}_k), \tag{3.2}$$

where $\widehat{\Theta}_k$ is an approximate inverse of $\nabla^2\ell_k(\widetilde{\theta}_k)$. The choice of $\widehat{\Theta}_k$ in the correction term crucial to the performance of the debiased estimator. In particuar, $\widehat{\Theta}_k$ must satisfy

$$\|I - \widehat{\Theta}_k\nabla^2\ell_k(\widetilde{\theta}_k)\|_\infty \lesssim \left(\frac{\log d}{n_k}\right)^{\frac{1}{2}}.$$

One possible approach to forming $\widehat{\Theta}_k$, as in van de Geer et al. (2014), is by nodewise regression on the weighted design matrix $\mathbf{X}_{\widehat{\theta}_k} := W_{\widehat{\theta}_k}\mathbf{X}_k$, where, $\mathbf{X}_k$ is the $n_k \times d$ design matrix for k-th dataset, defined as $\mathbf{X}_k = \begin{bmatrix}\mathbf{x}_{k1} & \mathbf{x}_{k2} & \dots & \mathbf{x}_{kn_k}\end{bmatrix}^\top$, and $W_{\widehat{\theta}_k}$ is $n_k \times n_k$ diagonal matrix, whose diagonal entries are

$$\left(W_{\widehat{\theta}_k}\right)_{i,i} := \ddot{\rho}(y_{ki}, x_{ki}^T\widehat{\theta}_k)^{\frac{1}{2}}.$$

That is, for $j \in [p]$ that machine k is debiasing, the machine solves

$$\widehat{\boldsymbol{\gamma}}_j^{(k)} := \arg\min_{\gamma\in\mathbf{R}^{p-1}}\frac{1}{2n_k}\|\mathbf{X}_{\widehat{\theta}_k,j} - \mathbf{X}_{\widehat{\theta}_k,-j}\gamma\|_2^2 + \lambda_j\|\gamma\|_1,\ j \in [p], \tag{3.3}$$

and forms

$$\left(\widehat{\Theta}_k\right)_{j,\cdot} = \frac{1}{\hat{\tau}_{kj}^2}\left[-\widehat{\boldsymbol{\gamma}}_{j,1}^{(k)} \quad \dots \quad -\widehat{\boldsymbol{\gamma}}_{j,j-1}^{(k)} \quad 1 \quad -\widehat{\boldsymbol{\gamma}}_{j,j+1}^{(k)} \quad \dots \quad -\widehat{\boldsymbol{\gamma}}_{j,p}^{(k)}\right], \tag{3.4}$$

where

$$\hat{\tau}_{kj} = \left(\frac{1}{n_k}\|\mathbf{X}_{\widehat{\theta}_k,j} - \mathbf{X}_{\widehat{\theta}_k,-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2 + \lambda_j\|\widehat{\boldsymbol{\gamma}}_j\|_1\right)^{\frac{1}{2}}.$$

After calculating the *debiased lasso estimators* $\widetilde{\theta}_k^d$ in local datasets, the integrated estimator is obtained using similar optimization problem as in (2.3). j-th co-ordinate of the *integrated debiased lasso* estimator is obtained as

$$\left(\widetilde{\theta}_0\right)_j = \arg\min_x\sum_{k=1}^m\Psi_{\eta_j}\left(\left(\widetilde{\theta}_k^d\right)_j - x\right), \tag{3.5}$$

where, η_j's used in the above equation are the same ones that are used to identify the global parameter. The integrated estimator $\tilde{\theta}_0$ calculated in (3.5) has a serious drawback. Loosely

speaking, the use of quadratic re-descending loss to define $\widetilde{\theta}_0$ gives us the co-ordinate wise mean of debiased lasso estimates, after dropping the outlier values. Since the debiased estimates are no longer sparse, $\widetilde{\theta}_0$ is also dense. This detracts from the interpretability of the coefficients and makes the estimation error large in the ℓ_2 and ℓ_1 norms. To remedy both problems, we threshold $\widetilde{\theta}_0$. Below we describe the hard-threshold and soft-threshold on $\widetilde{\theta}_0$:

$$HT_t\left(\widetilde{\theta}_0\right) \leftarrow \left(\widetilde{\theta}_0\right)_j \cdot \mathbf{1}_{\left\{\left|\left(\widetilde{\theta}_0\right)_j\right| \geq t\right\}}, \text{or} \tag{3.6}$$
$$ST_t\left(\widetilde{\theta}_0\right) \leftarrow \mathsf{sign}\left(\left(\widetilde{\theta}_0\right)_j\right) \cdot \max\left\{\left|\left(\widetilde{\theta}_0\right)_j\right| - t,\, 0\right\}.$$

The final estimator $\widehat{\theta}_0$ is obtained by hard-thresholding or soft-thresholding $\widetilde{\theta}_0$ at $t \sim \sqrt{\frac{\log d}{N_{\min}}}$, where, $N_{\min} = m \min_k n_k$, *i.e.*, $\widehat{\theta}_0 = HT_t\left(\widetilde{\theta}_0\right)$ or $ST_t\left(\widetilde{\theta}_0\right)$, and the final estimators for $\boldsymbol{\delta}_k$'s are obtained by hard-thresholding or soft-thresholding $\widetilde{\theta}_k^d - \widetilde{\theta}_0$ at a level $t' \sim \sqrt{\frac{\log d}{n_k}}$, *i.e.*, $\widehat{\delta}_k \leftarrow HT_{t'}\left(\widetilde{\delta}_k\right)$ or $ST_{t'}\left(\widetilde{\delta}_k\right)$.

A step by step process to obtain the global estimator $\widehat{\theta}_0$ for linear regression models is described in Algorithm 1. The Algorithm requires suitable choices of the parameters $\{\eta_j\}$, t and $\{t_k\}$. In our numerical studies we pick these choices via the cross-validation approach, described in Algorithm 2. If appropriate choices of η_j's are available from background knowledge then one can use those instead of picking them from cross-validation. In practice, cross-validating over $(\{\eta_j\}, t, \{t_k\}))$ might be computationally challenging since there are $d + m + 1$ many unknown parameters for m local datasets and d as the covariate dimension; and specifically d can be quite large. In our experiments we further simplify the choice by assuming η_j's have equal value η and $t_k = t\sqrt{N_{\min}/n_k}$. This simplifies the cross-validation parameters $(\{\eta_j\}, t, \{t_k\}))$ to just two parameters (η, t).

Algorithm 1: MrLasso$(\{\eta_j\}, t, \{t_k\})$

Input: $\{\mathcal{D}_k = (\{\mathbf{x}_{ki}, \mathbf{y}_{ki}\})_{i=1}^{n_k} : k \in [m]\}$, $\{\eta_j\}_{j\in[d]}$.
Output: $\left(\widehat{\theta}_0, \widehat{\delta}_1, \ldots, \widehat{\delta}_m\right)$.
for $k = 1, 2, \cdots, m$, **do**
  $\widetilde{\theta}_k \leftarrow \arg\min_{\theta\in\mathbf{R}^d} \frac{1}{2n_k} \sum_{i=1}^{n_k} \rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta) + \lambda_k\|\theta\|_1$
  $\widetilde{\theta}_k^d \leftarrow \widetilde{\theta}_k - \widehat{\Theta}_k \nabla\ell_k(\widetilde{\theta}_k)$, where, $\widehat{\Theta}_k$ are defined as in 3.4,
$\left(\widetilde{\theta}_0\right)_j \leftarrow \arg\min_x \sum_{k=1}^m \Psi_{\eta_j}\left(\left(\widetilde{\theta}_k^d\right)_j - x\right)$,
$\widehat{\theta}_0 \leftarrow HT_t\left(\widetilde{\theta}_0\right)$ or $ST_t\left(\widetilde{\theta}_0\right)$ for $t \sim \sqrt{\frac{\log d}{N_{\min}}}$, where, $N_{\min} = m \min_k n_k$,
for $k \in [m]$, **do**
  $\widetilde{\delta}_k \leftarrow \widetilde{\theta}_k^d - \widetilde{\theta}_0$,
  $\widehat{\delta}_k \leftarrow HT_{t_k}\left(\widetilde{\delta}_k\right)$ or $ST_{t_k}\left(\widetilde{\delta}_k\right)$ for $t_k \sim \sqrt{\frac{\log d}{n_k}}$.

Algorithm 2: Cross-validation

Input: $\mathcal{D} = \{\mathcal{D}_k = (\{\mathbf{x}_{ki}, \mathbf{y}_{ki}\})_{i=1}^{n_k} : k \in [m]\}$, C, G the grid of $(\{\eta_j\}, t, \{t_k\})$.

Output: $\left(\widehat{\theta_0}, \widehat{\delta}_1, \ldots, \widehat{\delta}_m\right)$.

$\{\mathcal{D}^{(c)}\}_{c=1}^{C} \leftarrow C$ equal partition of $\mathcal{D}$

for $(\{\eta_j\}, t, \{t_k\}) \in G$ **do**

  for $c = 1, \ldots, C$ **do**

    $\mathcal{D}^{(-c)} \leftarrow \mathcal{D} \backslash \mathcal{D}^{(c)}$

    $\left(\widehat{\theta_0}^{(c)}, \widehat{\delta}_1^{(c)}, \ldots, \widehat{\delta}_m^{(c)}\right) \leftarrow \text{MrLasso}\big(\mathcal{D}^{(-c)}, (\{\eta_j\}, t, \{t_k\})\big)$

    $\mathcal{E}_c \leftarrow \frac{1}{|\mathcal{D}^{(c)}|} \sum_{k=1}^{m} \sum_{(x,y) \in \mathcal{D}_k^{(c)}} \rho\{y, x^\top (\widehat{\theta_0}^{(c)} + \widehat{\delta}_k^{(c)})\}$

  $\mathcal{E}(\{\eta_j\}, t, \{t_k\}) \leftarrow \frac{1}{C} \sum_{c=1}^{C} \mathcal{E}_c$

$(\{\hat{\eta}_j\}, \hat{t}, \{\hat{t}_k\}) \leftarrow \arg\min \mathcal{E}(\{\eta_j\}, t, \{t_k\})$

$\left(\widehat{\theta_0}, \widehat{\delta}_1, \ldots, \widehat{\delta}_m\right) \leftarrow \text{MrLasso}\big(\mathcal{D}, (\{\hat{\eta}_j\}, \hat{t}, \{\hat{t}_k\})\big)$

4. Theoretical properties of the communication-efficient estimator

To present the theoretical justification of consistency of the estimators we shall focus on ℓ_1, ℓ_2, and ℓ_∞ consistency of the estimators. Before getting into the assumptions and results, we first define some notations. We use $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ to denote usual ℓ_1-norm, ℓ_2-norm, and ℓ_∞-norm respectively.

The performance of global estimator depends on the debiased lasso estimators from the local datasets. Hence, it is important to have a reasonable performance for local estimators. We study the ℓ_∞ error rate of the debiased lasso estimator $\widetilde{\theta}^d$ for the parameter θ, calculated form the dataset $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$. For a random vector $(\mathbf{y}, \mathbf{x})$, whose distribution is parametrized by θ, we assume that $\mathbb{E}[\rho(\mathbf{y}, \mathbf{x}^T\theta)]$ is uniquely minimized at θ. As before, the debiased lasso estimator is defined as $\widetilde{\theta}^d \triangleq \widetilde{\theta} - \widehat{\Theta}\nabla\ell(\widetilde{\theta})$, where, $\widetilde{\theta} = \arg\min_{\theta \in \mathbf{R}^d} \ell(\theta) + \lambda\|\theta\|_1$, for $\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \rho(\mathbf{y}_i, \mathbf{x}_i^T\theta)$, and $\widehat{\Theta}$ is calculated according to (3.4). We assume that the actual parameter value θ^* is s_0 sparse, *i.e.*, θ^* has s_0 many non-zero entries. Letting $\Theta \triangleq (\mathbb{E}\nabla^2\ell(\theta^*))^{-1}$ we assume the j-th row of Θ, which is denoted as $\Theta_{j,\cdot}$ is s_j-sparse. Denote $s^* = \max_j s_j$. We make the following assumptions on distributions of local datasets. These assumptions establish high probability bounds for the local debiased lasso estimates.

Assumption 1 *Under the M-estimation setup, we assume the following.*

1. *The pairs $\{(x_i, y_i)\}_{i=1}^n$ are* iid *P and for some $K \geq 1$, $\|\mathbf{X}\|_\infty = \max_{i,j}|X_{i,j}| = \mathcal{O}(K)$ and the projection of $\mathbf{X}_{\theta^*,j}$ on the row space of $\mathbf{X}_{\theta^*,-j}$ in the $\Sigma_\theta^* \triangleq \mathbb{E}[\nabla^2\ell(\theta^*)]$ inner product is bounded: $\|\mathbf{X}_{\theta^*,-j}\gamma_j\|_\infty = \mathcal{O}(K)$ for any j, where*

$$\gamma_j \triangleq \arg\min_{\gamma \in \mathbf{R}^{p-1}} \mathbb{E}\left[\|\mathbf{X}_{\theta^*,j} - \mathbf{X}_{\theta^*,-j}\gamma\|_2^2\right].$$

   We define τ_j^2 as $\tau_j^2 \triangleq \mathbb{E}\big[\frac{1}{n}\|\mathbf{X}_{\theta^*,j} - \mathbf{X}_{\theta^*,-j}\gamma_j\|_2^2\big]$.
2. *It holds that $K^2 s_j \sqrt{\log d/n} = o(1)$.*
3. *The smallest eigenvalue of Σ_{θ^*} is bounded away from zero, and moreover, $\|\Sigma_{\theta^*}\|_\infty = \mathcal{O}(1)$.*

4. *There exists a $\delta > 0$ such that for any θ that satisfies $\|\theta - \theta^*\|_1 \le \delta$ it holds that $\ddot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta)$ stay away form zero and that $\|\ddot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta)\|_\infty = \mathcal{O}(1)$. Furthermore, for any such θ and all x, y*
$$|\ddot{\rho}(y, x^T\theta) - \ddot{\rho}(y, x^T\theta^*)| \le |x^T(\theta - \theta^*)|.$$
5. *With probability at least $1 - o(d^{-1})$ we have $\frac{1}{n}\|\mathbf{X}(\widetilde{\theta} - \theta^*)\|_2^2 \lesssim s_0 \frac{\log d}{n}$ and $\|\widetilde{\theta} - \theta^*\|_1 \lesssim s_0\sqrt{\frac{\log d}{n}}$.*
6. *The derivative $\dot{\rho}(y, a)$, $\ddot{\rho}(y, a)$ exists for all y, a, and for some δ-neighborhood, $\ddot{\rho}(y, a)$ is locally Lipschitz: for some $L > 0$*
$$\max_{a_0 \in \{x_i^T\theta^*\}} \sup_{|a - a_0| \vee |\widehat{a} - a_0| \le \delta} \sup_{y \in \mathcal{Y}} \frac{|\ddot{\rho}(y, a) - \ddot{\rho}(y, \widehat{a})|}{|a - \widehat{a}|} \le L.$$
Moreover,
$$\max_{a_0 \in \{x_i^T\theta^*\}} \sup_{y \in \mathcal{Y}} |\dot{\rho}(y, a_0)| = \mathcal{O}(1), \quad \max_{a_0 \in \{x_i^T\theta^*\}} \sup_{|a - a_0| \le \delta} \sup_{y \in \mathcal{Y}} |\ddot{\rho}(y, a_0)| = \mathcal{O}(1).$$
7. *The diagonal entries of $\Theta\mathbb{E}\left[\nabla\ell(\theta^*)\nabla\ell(\theta^*)^T\right]\Theta$ are bounded by σ^2.*

The list of conditions in Assumption 1 are standard ones to establish ℓ_∞ convergence for the debiased lasso estimator and are adapted from (van de Geer et al., 2014, Assumptions (D1)-(D5)) and (Lee et al., 2017, Assumptions (B1)-(B7)). Condition 5 that directly assumes convergence of estimation and prediction error for lasso is implied by the other assumptions. We refer to (Bühlmann and Van De Geer, 2011, Chapter 6, Section 6.7) for the details, where the necessary compatibility condition is inherited from the condition 3. Here we state this to simplify the exposition. We refer to the corresponding literature (specifically Lee et al. (2017)) for further discussions about the conditions.

In the literature on debiased lasso for high-dimensional M-estimation, it is usually assumed that $K = \mathcal{O}(1)$ (see (van de Geer et al., 2014, Section 3.3.1), (Lee et al., 2017, Section 5)). For sub-gaussian covariates one can use a high-probability bound of the order $K = \mathcal{O}_{\mathbb{P}}(\sqrt{\log(dn)})$. The high probability ℓ_∞ bound for local debiased lasso estimate follows.

Lemma 2 *Under Assumption 1 the debiased lasso estimator $\widetilde{\theta}^d$ satisfies the following high probability bound.*
$$\text{for some } c > 0,\ \mathbb{P}\left(\|\widetilde{\theta}^d - \theta^*\|_\infty > \sigma\sqrt{\frac{12\log d}{n}} + c\frac{\max\{Ks^*, K^2 s_0\}\log d}{n}\right) \le o(d^{-1}). \quad (4.1)$$

Remark 3 *Under the Assumption 1 and $K^2 s_0\sqrt{\frac{\log d}{n}} = o(1)$ we have a high probability bound $\|\widetilde{\theta}^d - \theta^*\|_\infty \lesssim_{\mathbb{P}} \sqrt{\log d/n}$.*

Before studying the theoretical properties of the integrated estimator defined in (3.5) one might notice that the identification restriction (2.3) may not lead to a unique global parameter, in general. The whole business of studying the theoretical properties of the

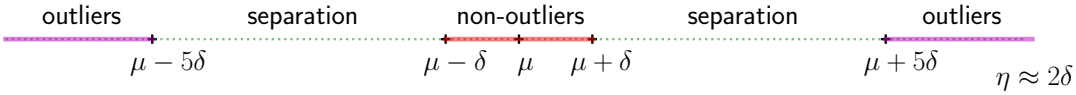

Figure 2: Bulk structure for the local parameters (Assumption 4). This structure ensures the global parameter (2.3) is unique.

global estimator is not meaningful if one cannot uniquely identify the global parameter. Hence, it is necessary to make some assumptions about the parameter values, with the goal that they will suffice unique identification of the global parameter. Here, we present the assumptions on the finite number of machines to have a uniquely identified parameter. For $j \leq d$ we assume

Assumption 4 *(B1) Let I_j be the set of indices for $(\theta_k^*)_j$'s which are considered as inliers. We assume $|I_j|/m \geq 4/7$.*

(B2) Let $\mu_j = \frac{1}{|I_j|}\sum_{k\in I_j}(\theta_k^*)_j$. Let δ be the smallest positive real number such that $(\theta_k^*)_j \in [\mu_j - \delta, \mu_j + \delta]$ for all $k \in I_j$. We assume that none of the $(\theta_k^*)_j$'s are in the intervals $[\mu_j - 5\delta, \mu_j - \delta)$ or $(\mu_j + \delta, \mu_j + 5\delta]$.

(B3) Let $\delta_2 = \min_{k_1\in I_j, k_2 \notin I_j} |(\theta_{k_1}^*)_j - (\theta_{k_2}^*)_j|$ is the minimum separation between inliers and outliers. Clearly, $4\delta < \delta_2$. We choose η_j such that $2\delta < \eta_j < \delta_2/2$.

The Assumption 4 sets a co-ordinate wise clustering assumption for the local parameters. A visualization summary of the cluster Assumption 4 is seen in Figure 2. The conditions ensure enough separation between the bulk and the outliers such that the re-descending loss (2.3) uniquely identifies the global parameter at mean of the bulk values, as suggested by Result 5.

Result 5 *Under the Assumption 4 the objective function $x \mapsto \sum_{k=1}^m \Psi_{\eta_j}((\theta_k^*)_j - x)$ has a unique minimizer $(\theta_0^*)_j = \mu_j$.*

Result 5 implies that under the Assumption 4 the global parameter θ_0^* is uniquely identified at the mean of the inlier parameter values. This has the following nice implications: (1) the global parameter is not affected by outlier values; (2) the estimator of the global parameter can effectively borrow strength across datasets with inlier parameter values. The Lemma 6, where we study the coordinate-wise convergence rates for $\widetilde{\theta}_0$, is a step toward establishing effective borrowing strength for the estimator of global parameter.

Lemma 6 *Let the following hold:*

(i) For any j, $\{(\theta_k^*)_j\}_{k=1}^n$ satisfy the Assumption 4.

(ii) The datasets $\{\mathcal{D}_k\}_{k=1}^m$ satisfy the Assumption 1 uniformly over k.

(iii) Let $s_{k,0}$ be the sparsity for θ_k^* and s_k^* being the maximum sparsity for the rows of $\Theta_k \triangleq \{\mathbb{E}[\nabla^2 \ell_k(\theta_k^*)]\}^{-1}$. Define $s_{0,\max} = \max_k s_{k,0}$, $s_{\max}^* = \max_k s_k^*$ and $s_{\text{eff}} = \max\{Ks_{\max}^*, K^2 s_{0,\max}\}$. We assume $m = o\left(\frac{n_{\min}^2}{s_{\text{eff}}^2 n_{\max}\log d}\right)$, where, $n_{\max} = \max_{k\in[m]} n_k$, and $n_{\min} = \min_{k\in[m]} n_k$.

Then for sufficiently large n_k we have the following bound for the co-ordinates of $\widetilde{\theta}_0$ and ℓ_∞ bound for $\widetilde{\delta}_k$:

$$\left|(\widetilde{\theta}_0)_j - (\theta_0^*)_j\right| \leq 4\sigma\sqrt{\frac{\log d}{|I_j|^2}\sum_{k\in I_j}\frac{1}{n_k}},\ \textit{for all } j,\ \textit{and } \|\widetilde{\delta}_k - \delta_k^*\|_\infty \leq 4\sigma\sqrt{\frac{\log d}{n_k}},\ \textit{for all } k, \tag{4.2}$$

with probability at least $1 - o(1)$.

Condition (i) in lemma 6 is needed to identify the global parameter uniquely, as suggested by Result 5. Condition (ii) provides a high probability ℓ_∞ bound for the local debiased lasso estimates (see lemma 2). Finally, the condition (iii) ensures the bias in local debiased estimates is of smaller order than the variance in integrated estimate. Since the bias term may not improve under averaging, violating this condition would result in higher-order for the bias in local estimates, and the estimation error for the global parameter stops improving (with a higher number of datasets).

In the following remark we introduce a weighted version of our data integration method (3.5).

Remark 7 (A weighted integration) *One may also consider* a weighted version *of the data integration*

$$\left(\widetilde{\theta}_0^w\right)_j = \arg\min_x \sum_{k=1}^m w_k \Psi_{\eta_j}\left(\left(\widetilde{\theta}_k^d\right)_j - x\right), \tag{4.3}$$

where $w_k \geq 0$ is the weight corresponding to k-th dataset with $\sum_{k=1}^m w_k = 1$. In that case the uniqueness of the corresponding global parameter

$$\left(\theta_0^{w,*}\right)_j = \arg\min_x \sum_{k=1}^m w_k \Psi_{\eta_j}\left((\theta_k^*)_j - x\right), \tag{4.4}$$

at the value $\left(\theta_0^{w,*}\right)_j = \sum_{k\in I_j} w_k(\theta_k^*)_j/(\sum_{k\in I_j} w_k)$ is established (in Result 17) by considering a weighted version of Assumption 4 (see Assumption 16) where we replace $|I_j|/m$ by $\sum_{k\in I_j} w_k$ in condition (B1), and let $\mu_j = \sum_{k\in I_j} w_k(\theta_k^*)_j/(\sum_{k\in I_j} w_k)$ in condition (B2). Furthermore, under conditions (ii) and (iii) in Lemma 6, one can establish (see Theorem 18) the following high probability bound for $\left(\widetilde{\theta}_0^w\right)_j$:

$$\mathbb{P}\left[\left|(\widetilde{\theta}_0^w)_j - (\theta_0^{w,*})_j\right| \leq 4\sqrt{\frac{\log d}{\left(\sum_{k\in I_j} w_k\right)^2}\sum_{k\in I_j}\frac{w_k^2}{n_k}},\ \textit{for all } j\right] = 1 - o(1)\,.$$

Further details about the setup are provided in Appendix C.

An interesting special case arises when $w_k = n_k/N$, and the above rate simplifies to $\mathcal{O}\big(\{\sum_{k\in I_j} n_k\}^{-1}\big)$ and yields the fastest rate among all possible weights:

$$\frac{1}{\sum_{k\in I_j} n_k} = \min_w \Big[\frac{1}{\left(\sum_{k\in I_j} w_k\right)^2}\sum_{k\in I_j}\frac{w_k^2}{n_k}\Big]\,.$$

Note that the rate $\mathcal{O}\big(\{\sum_{k\in I_j} n_k\}^{-1}\big)$ *is faster than the corresponding one in* (4.2)*, especially when the* n_k*'s are significantly different. The downside to the weighted formulation is that the corresponding structural assumptions on the local parameter values depends on the sample size, which can be perceived as strange.*

Proof Here we shall provide a brief outline of the proof of Lemma 6. The detailed proof of this lemma is provided in appendix B. We start with the Remark 3, which gives us a high probability error bound for the individual local lasso estimators. From condition (i) we see that the θ_j^*'s satisfies Assumption 4. As long as $\widetilde{\theta}_k^d$ concentrates around θ_k^*, $\widetilde{\theta}_k^d$ also satisfies Assumption 4 (with different δ_j and $\delta_{2,j}$'s but identical η_j's). Using Result 5 we see that the integrated estimator has the form

$$(\widetilde{\theta}_0)_j = \frac{\sum_{k\in I_j}(\widetilde{\theta}_k^d)_j}{|I_j|}.$$

Finally, we use the decomposition on the individual lasso estimators $\tilde{\theta}_k^d$ in Lemma 12 and concentration bound for sums of independent random variables to get a high probability bound for the estimation error of $\widetilde{\theta}_0$. ■

Before we study the theoretical properties of the thresholded estimators, we would like to make a few remarks about the rates of convergence for the coordinates of $\widetilde{\theta}_0$. If there is reason to believe that none of the local parameters are outliers and we consider η_j's large enough, then the global parameter is identified as mean of the local parameters, *i.e.* $\theta_0^* = \frac{1}{m}\sum_{k=1}^m \theta_k^*$. In that case, the rate of convergence for $\widetilde{\theta}_0$ in Lemma 6 simplifies to:

$$\|\widetilde{\theta}_0 - \theta_0^*\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{m^2}\sum_{k=1}^m \frac{1}{n_k}},$$

where, $\lesssim_{\mathbb{P}}$ implies with probability converging to 1. This kind of rate is not surprising. Probably the simplest situation, under which this kind of convergence rate arises is ANOVA model. Let $\{Y_{ki}\}_{i\in[n_k],k\in[m]}$ is a set of independent random variables, such that $Y_{ki} \sim N(\theta_0^* + \delta_k^*, \sigma^2)$, where $\theta_0^*, \delta_1^*, \ldots, \delta_m^*$ are some real numbers. Under the assumption $\sum_{k=1}^m \delta_k^* = 0$, the estimator $\widehat{\theta}_0 = \frac{1}{m}\sum_{k=1}^m \bar{Y}_{k\cdot}$, where, $\bar{Y}_{k\cdot} = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki}$, follows distribution $N\left(\theta_0^*, \frac{1}{m^2}\sum_{k=1}^m \frac{1}{n_k}\right)$. This gives us the following convergence rate for $\widehat{\theta}_0$:

$$\mathbb{P}\left(|\widehat{\theta}_0 - \theta_0^*| \le 3\sqrt{\frac{1}{m^2}\sum_{k=1}^m \frac{1}{n_k}}\right) \approx 1.$$

We notice that $\frac{1}{\sum_{k\in I_j} n_k} \le \frac{1}{|I_j|^2}\sum_{k\in I_j}\frac{1}{n_k}$, and the equality holds when n_k's are equal, which gives us the best possible rate. In this case the rate simplifies to:

$$\left|(\widetilde{\theta}_0)_j - (\theta_0^*)_j\right| \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{|I_j| n}} \le \sqrt{\frac{\log d}{(1-\alpha_j)N}}.$$

Let $n_{\min} = \min_k n_k$ and $\alpha_{\max} = \max_j \alpha_j$. Notice that $\frac{1}{\sum_{k\in I_j} n_k} \leq \frac{1}{|I_j|^2}\sum_{k\in I_j}\frac{1}{n_k} \leq \frac{1}{|I_j|n_{\min}} \leq \frac{1}{(1-\alpha_j)mn_{\min}}$. Hence, we can get the following simple (possibly naive) ℓ_∞ high probability bound for $\widetilde{\theta}_0$:

$$\|\widetilde{\theta}_0 - \theta_0^*\|_\infty \leq 4\sigma\sqrt{\frac{\log d}{(1-\alpha_{\max})mn_{\min}}}.$$

The estimators in Lemma 6 are dense, they have higher ℓ_1 and ℓ_2 error. We threshold the estimators at a suitable level. The thresholding is usually done at a level of the ℓ_∞ error rate of the estimator. We recall the following result from Lee et al. (2017) which ensures probability convergence for such thresholding.

Lemma 8 (Lee et al. (2017), Lemma 11) *As long as $t > \|\bar{\beta} - \beta^*\|_\infty$ $\bar{\beta}_t^{ht} = \mathrm{HT}_t(\bar{\beta})$ satisfies the following.*

1. $\|\bar{\beta}^{ht} - \beta^*\|_\infty < 2t$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_2 < 2\sqrt{2s}t$,
3. $\|\bar{\beta}^{ht} - \beta^*\|_2 < 2\sqrt{2}st$,

where, s is the sparsity of β^*. The analogous results holds for $\bar{\beta}_t^{st} = \mathrm{ST}_t(\bar{\beta})$.

A combination of the Lemmas 6 and 8 establishes the rate of convergence for the estimates of global and heterogeneous effect parameters.

Theorem 9 *Define $N_{\min} := mn_{\min}$. Assume that the threshold for $\widetilde{\theta}_0$ is set at $t_0 = 4\sigma\sqrt{\frac{\log d}{(1-\alpha_{\max})N_{\min}}}$ and for $\widetilde{\delta}_k$'s are set at $t_k = 4\sigma\sqrt{\log d/n_k}$, respectively. Then under the conditions in Lemma 6 and for sufficiently large n_k we have the following:*

1. $\|\widehat{\theta}_0 - \theta_0^*\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{N_{\min}}}$,
2. $\|\widehat{\theta}_0 - \theta_0^*\|_1 \lesssim_{\mathbb{P}} s(\theta_0^*)\sqrt{\frac{\log d}{N_{\min}}}$,
3. $\|\widehat{\theta}_0 - \theta_0^*\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{s(\theta_0^*)\log d}{N_{\min}}}$,
4. $\|\widehat{\delta}_k - \delta_k^*\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{n_k}}$,
5. $\|\widehat{\delta}_k - \delta_k^*\|_1 \lesssim_{\mathbb{P}} s(\delta_k^*)\sqrt{\frac{\log d}{n_k}}$,
6. $\|\widehat{\delta}_k - \delta_k^*\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{s(\delta_k^*)\log d}{n_k}}$,

where, for a vector $\theta \in \mathbf{R}^d$, $s(\theta)$ denotes its sparsity level.

It is not necessary to threshold all the co-ordinates of $\widetilde{\theta}_0$ at a same level. Thresholding the j-th co-ordinate of $\widetilde{\theta}_0$ at $t_{0j} = 4\sigma\sqrt{\frac{\log d}{|I_j|^2}\sum_{k\in I_j}\frac{1}{n_k}}$ may give us a faster rate of convergence for $\widehat{\theta}_0$ in terms of ℓ_1 and ℓ_2 errors. If one wish to use cross-validation to determine an appropriate choices of thresholding, setting different thresholding level for different co-ordinates may be computationally hectic. For computational simplicity, we consider same thresholding level for all the co-ordinates.

One important consequence of Theorem 9 is variable selection under beta-min assumption. For a vector $\theta \in \mathbf{R}^d$ let us define its sparsity set as $\mathcal{S}(\theta) = \{j \in [d] : \theta_j \neq 0\}$. Under the assumption that

$$\min_{j \in \mathcal{S}(\theta_0^*)} |\theta_j^*| \gg \sqrt{\frac{\log d}{N_{\min}}} \text{ and, } \min_{j \in \mathcal{S}(\theta_k^*)} |(\theta_k^*)_j| \gg \sqrt{\frac{\log d}{n_k}} \tag{4.5}$$

Theorem 9 implies $\mathbb{P}\left(\mathcal{S}(\widehat{\theta}_0) = \mathcal{S}(\theta_0^*), \mathcal{S}(\widehat{\theta}_k) = \mathcal{S}(\theta_k^*)\right) \to 1$, *i.e.,* with very high probability all the active variables will be selected by the proposed estimators.

5. Computational results

In this section, we compare the performance of MrLasso to several other global parameters considered in Section 2 on simulated data: the Mean (1.2), the Median (2.1), the square and absolute error trade-off in (2.5) and the Huber estimator in (2.2). The datasets are generated from a linear model with $d = 2000$ covariates. In our simulation, the covariates are generated from auto-regressive model ($x = (x_1, \ldots, x_d)^\top$, $x_1 \sim N(0,1)$, $x_j = \sqrt{1-\rho^2}x_{j-1} + \rho\epsilon_j$, and $\epsilon_j \sim N(0,1)$) with the correlation $\rho = 0.9$, and the response in the k-th local dataset is generated as

$$y_k = \mathbf{x}_k^\top \beta_k^* + \epsilon_k,\ \epsilon_k \sim N(0, 0.25),$$

where β_k^* is the vector of regression coefficients for the k-th dataset. The first s co-ordinates of β_k^*'s are independent $N(2, (1.5)^2)$ random scalars. The next 20 coordinates of β_k^* are drawn from the mixture distribution $(1 - \frac{1}{m}) \cdot \delta_0 + \frac{1}{m} \cdot U([15, 16])$, where δ_0 is the degenerate distribution at zero. This endows the local coordinates with the cluster structure of Assumption 4. This structure in the local coordinates highlights the difference between the identification restrictions in MrLasso and the other estimators.

Note that the definition MrLasso, square and absolute loss trade-off and Huber loss require additional tuning parameters. In the simulation, we hold out $\frac{1}{5}$ of the dataset as a validation set and we pick those parameters to minimize test error on the validation set. The data dependent choice of tuning parameters are used to define the global parameters for square and absolute loss trade-off and Huber loss to keep the comparison fair, whereas for MrLasso the global parameters are identified with oracle tuning parameter values. The upper left plot in Figure 3 shows the ℓ_2-estimation error of the global parameter with respect to the sample size for each dataset (n_k). We fix the number of datasets (m) at 5 and the sparsity of the MrLasso global parameter (s) at 5. Note that the global parameters are different for different identification restrictions. The convergence rate of MrLasso is faster than the others. For denser global parameters, the performance discrepancy between the estimators reduces. We see this behavior in the lower-left plot of Figure 3.

A study on the effect of m for fixed n_k and s is presented in the upper right plot of Figure 3. All the estimates are able to borrow strength across datasets (ℓ_2-error decreases for higher m) but MrLasso produces lower ℓ_2-errors than others.

In the lower right panel of Figure 3 we study the sensitivity of MrLasso to the misspecification of η. As seen in the plot, MrLasso produces a good estimate of the global parameter as long as η falls within a reasonable range. This range depends on the separation between the bulk and outlier local coordinates (see Assumption 4, condition (B3)). The plot also

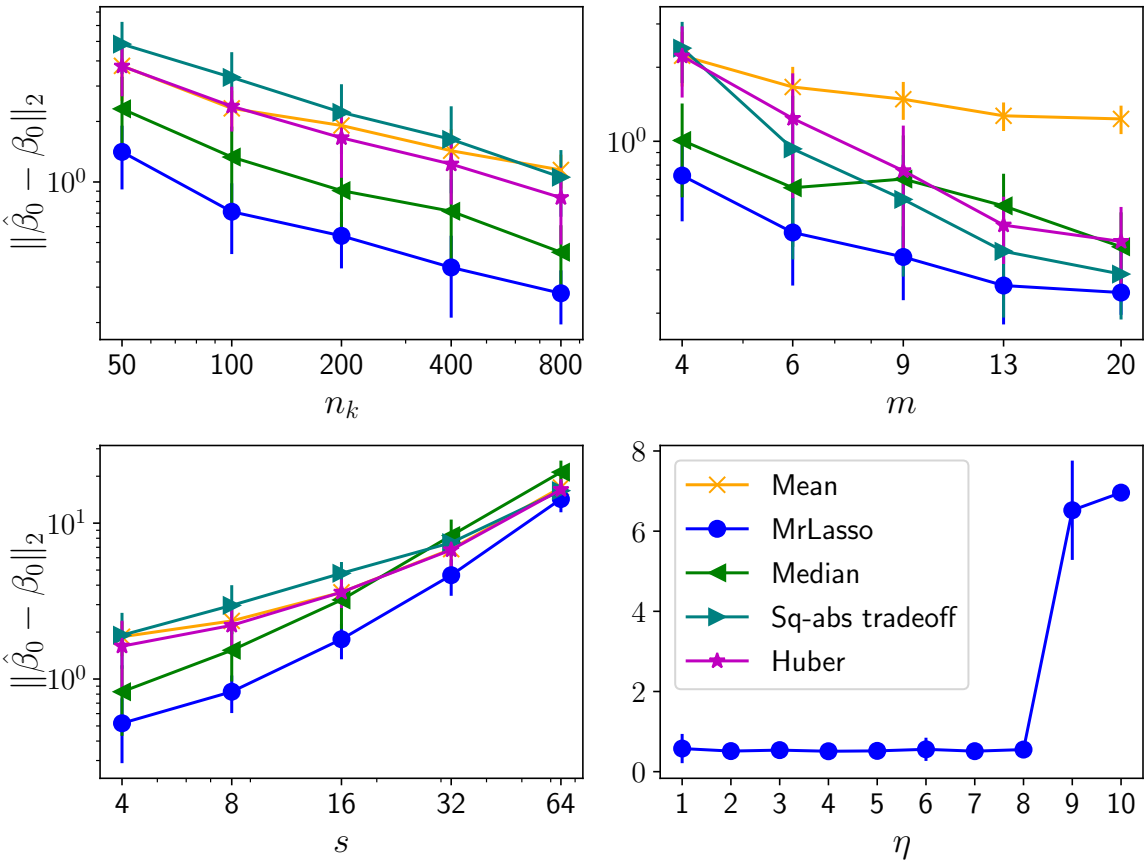

Figure 3: *Upper left*: Errorbar plots for ℓ_2-error in estimation of β_0 using different global estimators for varying sample sizes in each datasets (n_k) with $m = 5$ and $s = 5$. *Upper right*: For varying m with $n_k = 200$ and $s = 5$. *Lower left*: For different s with $n_k = 200$ and $m = 5$. *Lower right*: For different η with $n_k = 200$ and $s = 5$ and $m = 5$.

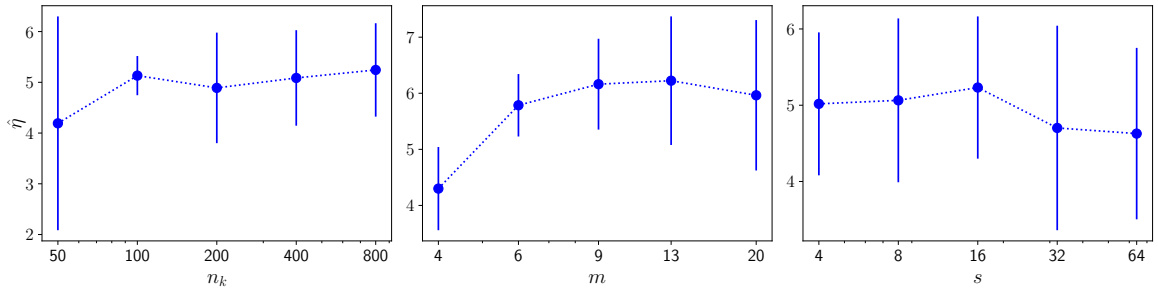

Figure 4: The choices of data dependent η's for MrLasso in the experiments in Figure 3, upper-left, upper-right and lower-left plot.

suggests that (cross-)validation leads to a value of η that falls in this range. We confirm this in Figure 4, which shows the chosen values of η for the simulations in Figure 3.

The ℓ_2 errors in Figure 3 are calculated in terms of the corresponding global parameters of the estimators, but the global parameters are different. We complement this experiment by evaluating the prediction error of them on a test dataset that is generated without any outlier local regression coefficients. In particular, the first s coordinates of the local regression coefficient vectors are generated independently from $N(2, (1.5)^2)$, while the rest of the coordinates are zero. In this setup, we expect the MrLasso global parameter to be closer to the population global parameter because it correctly identifies the coordinates of all but the first s coordinates as zero. This is confirmed in Figure 5.

We finally note that our method is both privacy-preserving (no raw data is sent) and communication efficient (only scalars and vectors are sent). Even the model selection protocol for η can be carried out without having to send the raw data: we only need to transfer scalars and vectors. We first send the local debiased estimates to the central machine to calculate $\hat{\beta}_0$ and $\hat{\delta}_k$'s for several choices of η's. These estimates are sent back to the local machines to calculate prediction error on the holdout datasets. Finally these errors can be transferred back to the global machine to determine the choice of η.

6. Cancer cell line study

The Cancer Cell Line Encyclopedia is a database of gene expression, genotype, and drug sensitivity data for human cancer cell lines. We use our method to study the sensitivity of cancer cell lines to certain anti-cancer drugs. We treat the area above the dose-response curve as the response variable (Barretina et al., 2012) and use expression levels of approximately 20000 genes (d) as features. More details about data source and pre-processing are provided in Appendix E. After prepossessing we have 482 cancer cell lines. Each of these cell lines comes from a cancerous organ and corresponds to a specific cancer type. The lines are divided into different machines (k) according to cancer types. For a given drug, our goal is to obtain a single regression coefficient vector which can be used to predict the response of different cancer types to that drug. We fit our model on the more common cancer cell types [2] with at least 10 samples and evaluate the prediction accuracy of the models on rare cancer

2. The number of local datasets (m) that are used in meta-analysis is same as the number of common cancer cell types shown in the upper left panel of the corresponding figures (for example the drug PD-0325901

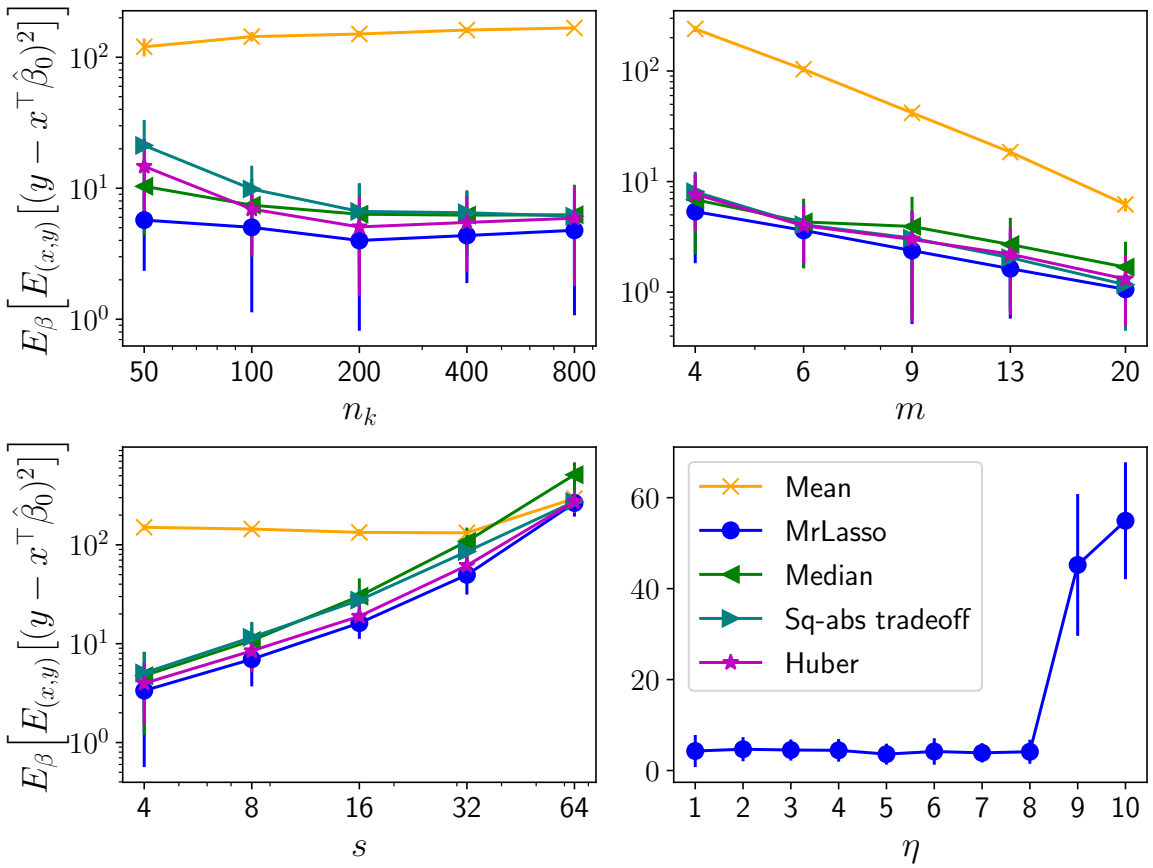

Figure 5: *Upper left*: Errorbar plots for prediction errors in test data using global estimators for varying sample sizes in each datasets (n_k) with $m = 5$ and $s = 5$. *Upper right*: For varying m with $n_k = 200$ and $s = 5$. *Lower left*: For different s with $n_k = 200$ and $m = 5$. *Lower right*: For different η with $n_k = 200$ and $s = 5$ and $m = 5$.

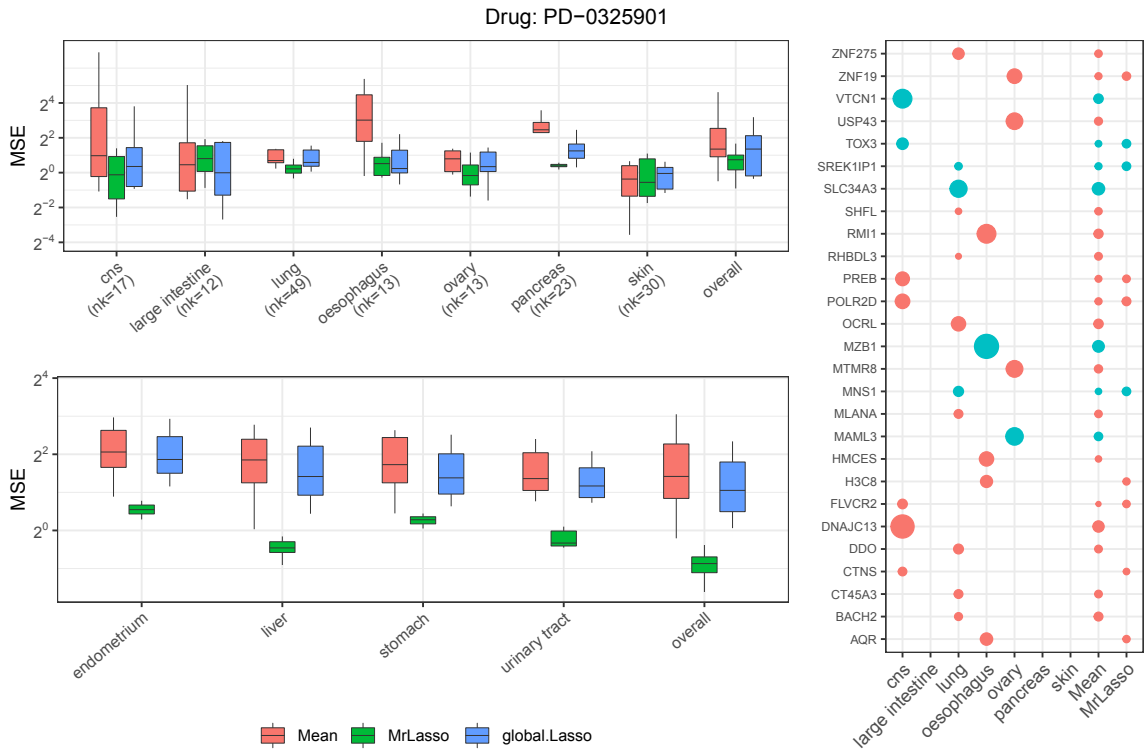

Figure 6: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug PD-0325901. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types (calculated locally on each machine) and Mean and Mr Lasso (integrated across machines). The size of the points is related to the magnitude of the coefficients, whereas the signs are represented by colour (red for negative and blue for positive).

types (those with less than 10 samples). The performance of three models are presented: soft thresholded mean of the local estimators (which we call Mean), the soft thresholded version of our integrated estimator (as in section 3), and a global lasso estimator that simply fits the lasso to an aggregate dataset consisting of samples from all common cancer types.

For integration purposes we first need to calculate the de-biased lasso estimate from each of these frequent cancer types. The regularization parameter for each of these cancer types is chosen using 7 fold cross-validation. After calculating the debiased lasso from each dataset, Mean is obtained by soft-thresholding their average, with the level of thresholding determined via 10 fold cross-validation. To calculate MrLasso, we assume that the parameters η_j are identical over the gene expressions. Besides determining this common parameter η we also require a proper soft-threshold level (t) for the integrated estimator. To this end the appropriate pair (η, t) is again chosen via cross validation (see the discussion in last paragraph of Section 3). We compare the performances of the Mean, the MrLasso and the

has $m = 7$ as seen in Figure 6), where the number of data-points (n_k's) in each of these common cancer types are also shown in the upper-left panel.

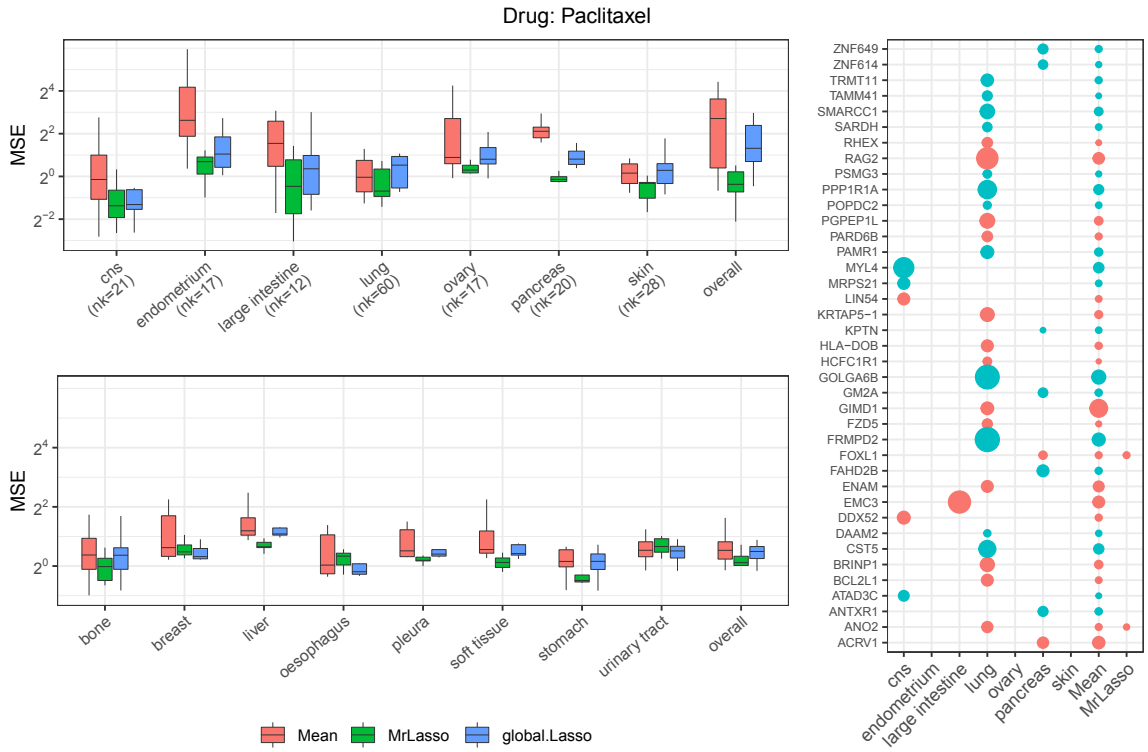

Figure 7: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug Paclitaxel. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and Mean and Mr Lasso.

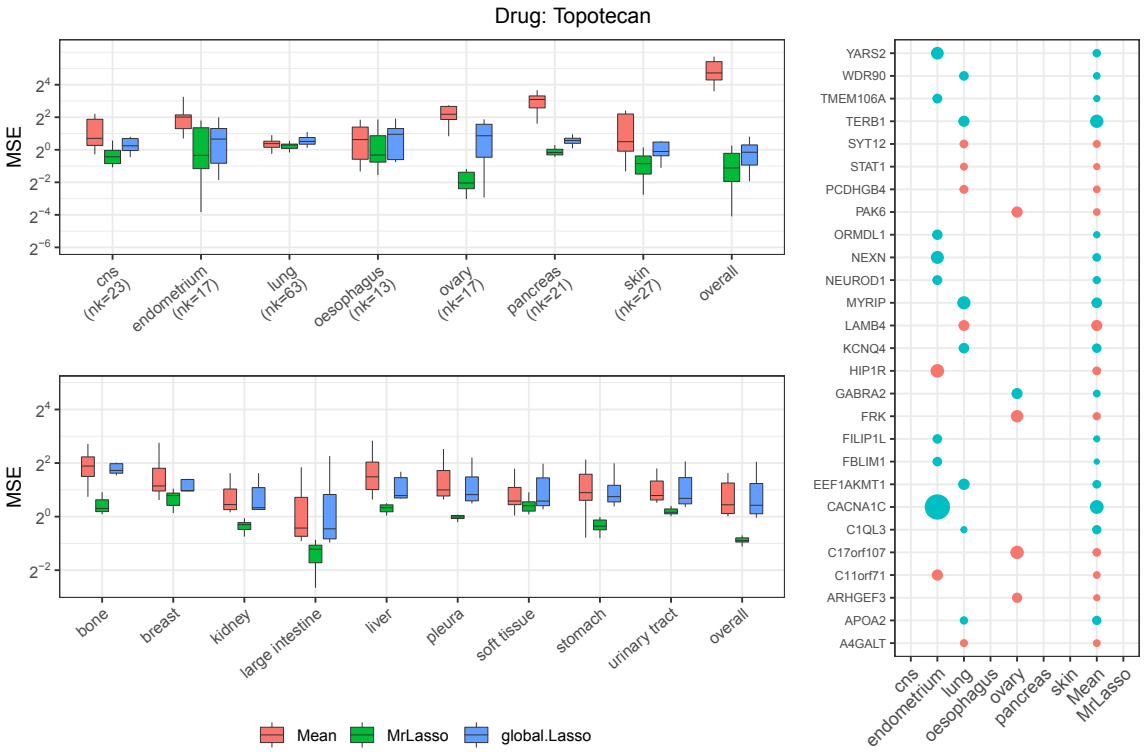

Figure 8: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug Topotecan. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and Mean and Mr Lasso.

global lasso (obtained using combined data over the most-frequent cancer types with best regularization parameter chosen via 10-fold cross-validation) for drugs PD-0325901 (Figure 6), Paclitaxel (Figure 7) and Topotecan (Figure 8).

To evaluate the efficacy of this approach, we fit the global parameter using Mean, MrLasso, and a global lasso on the more common cancer types and evaluate its predictive accuracy on rare cancer types. The lower left panels in Figures 6, 7 and 8 give comparative plots for the prediction accuracy of drug-response of the three global parameter estimates (for the three drugs). The plots for several other drugs are provided in Appendix E. For completeness, we also evaluate the predictive performance of the fitted global parameters on held out data from the more common cancer types in the top left panels.

We see that the predictive performance of MrLasso is generally superior to that of Mean and the global lasso. This is mostly due to the heterogeneity in the relationship between drug sensitivity and genotype on different cancer types, as evinced by the considerable variation between the effect sizes of genetic variants on drug sensitivity from cancer to cancer: see for example, Figures 6, 7 and 8. This causes the global parameter fitted by the Mean to be considerably denser than the global parameter fitted by MrLasso. In other words, the Mean pools effect sizes across all cancer types regardless of whether the effect sizes are similar. This causes the global parameter fitted by the Mean to pick up many small effect sizes that are specific to certain cancer types, thereby detracting from the generalizability of the global parameter to other (rare) cancer types. On the other hand, MrLasso only aggregates effect sizes when the data suggests they are similar. In other words, the global parameter fitted by MrLasso only incorporates effects that *persist* across cancer types. This leads to a parameter whose prediction performance generalizes across cancer types. In some extreme cases (*e.g.* Figure 8), the effect sizes are so dissimilar that MrLasso does not borrow strength at all, but the Mean continues to pool the effect sizes. This leads to overall poor prediction performance. In comparison with Mean, we see that MrLasso can *automatically adapt* to the intrinsic level of heterogeneity across datasets, leading to a better generalization.

7. Discussion

We consider integrative regression in the high-dimensional setting with heterogeneous data sources. The two main issues that we address are (i) identifiability of the global parameter, and (ii) statistically and computationally efficient estimator of the global parameter. In many prior works on integrative regression in high dimensions, there is either no global parameter or a global parameter that is not properly identified (see Section 2 for a discussion of global parameters in prior work).

We suggest a way to identify the global parameter that addresses some of the drawback of prior approaches by appealing to ideas from robust location estimation. The main benefit of our suggestion is it is possible to estimate the global parameter at a rate that depends on the size of the combined dataset. We also proposed a statistically and computationally efficient estimator of the global parameter. By statistically efficient, we mean the estimation error vanishes at a rate that depends on the size of the combined datasets. By computationally efficient, we mean the communication cost of evaluating the estimator depends only on the dimension of the problem and the number of machines. Further, because no individual

samples are communicated between machines, evaluating the proposed estimator does not compromise the privacy of the individuals in the data sources.

References

A. Asiaee, S. Oymak, K. R. Coombes, and A. Banerjee. High Dimensional Data Enrichment: Interpretable, Fast, and Data-Efficient. *arXiv:1806.04047 [cs, stat]*, June 2018.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T. Cai, M. Liu, and Y. Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, pages 1–15, 2021.

X. Cheng, W. Lu, and M. Liu. Identification of homogeneous and heterogeneous variables in pooled cohort studies: Identification of Heterogeneity in Pooled Studies. *Biometrics*, 71 (2):397–403, June 2015. ISSN 0006341X. doi: 10.1111/biom.12285.

C. C. L. E. Consortium, G. of Drug Sensitivity in Cancer Consortium, et al. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84, 2015.

S. M. Corsello, R. T. Nagari, R. D. Spangler, J. Rossen, M. Kocak, J. G. Bryan, R. Humeidi, D. Peck, X. Wu, A. A. Tang, et al. Non-oncology drugs are a source of previously unappreciated anti-cancer activity. *BioRxiv*, page 730119, 2019.

B. DepMap. DepMap 21Q2 Public, May 2021.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7 (3):177–188, 1986.

S. M. Gross and R. Tibshirani. Data Shared Lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235, Sept. 2016. ISSN 01679473. doi: 10.1016/j.csda.2016.02.015.

F. R. Hampel, editor. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, digital print edition, 2005. ISBN 978-0-471-73577-9. OCLC: 255133771.

L. V. Hedges and I. Olkin. *Statistical methods for meta-analysis*. Academic press, 2014.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, Mar. 1964. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177703732.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-Efficient Distributed Statistical Inference. *arXiv:1605.07689 [cs, math, stat]*, May 2016.

J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor. Communication-efficient sparse regression: A one-shot approach. *Journal of Machine Learning Research*, 18, Jan. 2017.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, Oct. 2010. ISSN 1471-0064. doi: 10.1038/nrg2825.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, June 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1221.

Appendix A. Identifiability

Lemma 10 (Contaminated normal model) *Let*

$$x \sim (1-\epsilon)\mathcal{N}\big((\mu_0, \sigma^2\big) + \epsilon g,$$

where $\epsilon < 1/2$ and g is continuous density of contamination distribution. The expected redescending loss $L(\mu) = E[(x-\mu)^2 \wedge \eta^2]$ is minimized at μ_0 if

$$\int_{\mu_0-\eta}^{\mu_0+\eta} (x-\mu_0) g(x) dx = 0$$

and

$$(1-\epsilon)\left\{\Phi\Big(\frac{\eta}{\sigma}\Big) - \Phi\Big(-\frac{\eta}{\sigma}\Big) - \frac{2\eta}{\sigma}\phi\Big(\frac{\eta}{\sigma}\Big)\right\} + \epsilon\left\{G(\mu+\eta) - G(\mu-\eta) - 2\eta\big(g(\mu+\eta)+g(\mu-\eta)\big)\right\} > 0,$$

where ϕ and Φ are density and distribution functions of standard normal distribution.

Proof We note that

$$\begin{aligned} L'(\mu)/2 &= E\big[(\mu - x)\mathbb{1}(|x-\mu| \le \eta)\big] \\ &= (1-\epsilon) E_{(1/\sigma)\phi\big((x-\mu)/\sigma\big)}\big[(\mu - x)\mathbb{1}(|x-\mu| \le \eta)\big] \\ &\quad + \epsilon E_g\big[(\mu - x)\mathbb{1}(|x-\mu| \le \eta)\big] \end{aligned}$$

Here

$$\begin{aligned} &E_g\big[(\mu_0 + \delta - x)\mathbb{1}(|x-\mu_0-\delta| \le \eta)\big] \\ &= \delta \int_{\mu_0+\delta-\eta}^{\mu_0+\delta+\eta} g(x)dx - \int_{\mu_0+\delta-\eta}^{\mu_0+\delta+\eta} (x-\mu_0) g(x) dx \\ &= \delta\Big[G(\mu_0+\delta+\eta) - G(\mu_0+\delta-\eta)\Big] - \Big[\Gamma(\mu_0+\delta+\eta) - \Gamma(\mu_0+\delta-\eta)\Big], \end{aligned}$$

where $\Gamma'(x) = (x-\mu_0)g(x)$. From the first order Taylor expansion

$$\begin{aligned} &E_g\big[(\mu_0 + \delta - x)\mathbb{1}(|x-\mu_0-\delta| \le \eta)\big] \\ &\approx \delta\Big[G(\mu_0+\eta) + \delta g(\mu_0+\eta) - G(\mu_0-\eta) - \delta g(\mu_0-\eta)\Big] \\ &\quad - \Big[\Gamma(\mu_0+\eta) + \delta\eta g(\mu_0+\eta) - \Gamma(\mu_0-\eta) + \delta\eta g(\mu_0-\eta)\Big] \\ &\approx \delta\Big[G(\mu_0+\eta) - G(\mu_0-\eta) - \eta\big(g(\mu_0+\eta) + g(\mu_0-\eta)\big)\Big] - \Big[\Gamma(\mu_0+\eta) - \Gamma(\mu_0-\eta)\Big] \end{aligned}$$

If g is the density of $N(\mu_0, \sigma^2)$ then

$$\begin{aligned} &E_g\big[(\mu_0 + \delta - x)\mathbb{1}(|x-\mu_0-\delta| \le \eta)\big] \\ &\approx \delta\left[\Phi\Big(\frac{\eta}{\sigma}\Big) - \Phi\Big(-\frac{\eta}{\sigma}\Big) - \frac{2\eta}{\sigma}\phi\Big(\frac{\eta}{\sigma}\Big)\right]. \end{aligned}$$

From the requirement that

$$L'(\mu_0+\delta)=\begin{cases}0 & \text{if } \delta=0,\\ >0 & \text{if } \delta>0,\\ <0 & \text{if } \delta<0,\end{cases}$$

for any arbitrarily small δ, we have the lemma.

■

Example 1 *Consider an one dimensional case where we have $2m+1$ datasets each of size n. The local parameters of these datasets take the values $\mu_k^*=2k$ for $k=1,2,\cdots,2m+1$. In this case the global parameter μ_0^* will be identified as $2(m+1)$. Suppose we have some estimators for the local parameters $\widehat{\mu}_k$ which can be written as $\widehat{\mu}_k=\mu_k^*+\eta_k$ and for simplicity we assume η_k are normal random variables with mean zero standard deviation $\frac{\sigma}{\sqrt{n}}$. The example can also be extended for sub-Gaussian random variables.*

The parameter σ depends on the variances of observations. For simplicity we assume that σ is 1. For some $0<\delta<1$ if sample sizes $n\geq 8\log\left(\frac{4m+2}{\delta}\right)$ for each datasets then by union bound over sub-Gaussian inequality we get

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mu}_k-\mu_k^*|\leq\frac{1}{2}\ \textit{for all}\ k=1,2,\cdots,2m+1\right)&\geq 1-\sum_{k=1}^{2m+1}\mathbb{P}\left(|\eta_k|>t\right)\\ &\geq 1-2(2m+1)e^{-\frac{n}{8}}\\ &\geq 1-\delta.\end{aligned}$$

We can see that with probability at least $1-\delta$ the local estimators are ordered as $\widehat{\mu}_1<\widehat{\mu}_2<\cdots<\widehat{\mu}_{2m+1}$. Let this event be E. A natural choice for $\widehat{\mu}_0$ is the median of $\widehat{\mu}_k$'s under the identification condition (2.1). Clearly, under the event E, $\widehat{\mu}_0$ is $\widehat{\mu}_{m+1}$ and hence we have $\widehat{\mu}_0-\mu_0^*=\widehat{\mu}_{m+1}-\mu_{m+1}^*=\eta_{m+1}$. As η_{m+1} is $N(0,1/n)$, we see that

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mu}_0-\mu_0^*|>\frac{C}{\sqrt{mn}}\right)&\geq\mathbb{P}\left(|\eta_{m+1}|>\frac{C}{\sqrt{mn}}\right)-\mathbb{P}(E^c)\\ &\geq 2-2\Phi\left(\frac{C}{\sqrt{m}}\right)-\delta.\end{aligned}$$

For a fixed $C>0$ the above probability is larger than $1-\Phi\left(\frac{C}{\sqrt{m}}\right)$ whenever $n\geq 8\log\left(\frac{4m+2}{1-\Phi\left(\frac{C}{\sqrt{m}}\right)}\right)$. This is a contradiction to the fact that rate of convergence for $\widehat{\mu}_0$ to μ_0^* is $\frac{1}{\sqrt{mn}}$.

Example 2 *Let $\mu_1=\cdots=\mu_{m-1}=0$, and $\mu_m>0$. For any choice of $\lambda>0$, μ_0 as defined in (2.2) is never zero.*

Proof Define

$$L(x,\lambda)=\sum_{k=1}^{m}L_\lambda(x-\mu_k)=(m-1)L_\lambda(x)+L(\mu_m-x).$$

Since

$$L_\lambda(x)=\begin{cases}\frac{1}{2}x^2 & \text{if } |x|\le\lambda\\ \lambda\left(|x|-\frac{1}{2}\lambda\right) & \text{otherwise,}\end{cases}$$

we note that

$$\partial_x L_\lambda(x)=\begin{cases}x & \text{if } |x|\le\lambda\\ \lambda\cdot\text{sign}(x) & \text{otherwise,}\end{cases}$$

for $x\neq 0$ and $\partial_x L_\lambda(0)=0$, we have

$$\partial_x L(0,\lambda)=(m-1)\partial_x L_\lambda(0)+\partial_x L(\mu_m)=\lambda\cdot\text{sign}(\mu_m)\,. \tag{A.1}$$

Note that $\partial_x L(0,\lambda)$ is can never be zero, which concludes that the minimum is never achieved at zero.

■

Result 11 *Suppose the covariate dimension d is fixed. Assume*

1. $n_1=\cdots=n_m=n$,
2. *Var*$(\boldsymbol{\epsilon}_{ki})$ *are same over different k's,*
3. $\Sigma_k=\mathbb{E}\left(\mathbf{x}_{ki}\mathbf{x}_{ki}^T\right)$*'s are invertible,*
4. $\sum_{k=1}^m\|\theta_k^*-\theta_0\|_1+c\|\theta_0\|_1$ *has a unique minimizer* θ_0^*.

Define

$$(\widehat{\theta}_0,\widehat{\delta}_1,\ldots,\widehat{\delta}_m)(n)=\arg\min_{(\theta_0,\delta_1,\ldots,\delta_m)}\frac{1}{2N}\sum_{k=1}^{m}\|\mathbf{Y}_k-\mathbf{X}_k(\theta_0+\delta_k)\|_2^2+\lambda_n\sum_{k=1}^{m}\|\delta_k\|_1+c\lambda_n\|\theta_0\|_1,$$

for $\lambda_n\sim\frac{1}{\sqrt{n}}$ *some* $c>0$ *independent of* n. *Then*

$$(\widehat{\theta}_0,\widehat{\delta}_1,\ldots,\widehat{\delta}_m)(n)\to(\theta_0^*,\delta_1^*,\ldots,\delta_m^*),$$

in probability, with respect to $\|\cdot\|_2$ *norm, where,* $\theta_0^*=\arg\min_{\theta_0}\left(\sum_{k=1}^m\|\theta_k^*-\theta_0\|_1+c\|\theta_0\|_1\right)$ *and* $\delta_k^*=\theta_k^*-\theta_0^*$ *for* $k\in[m]$.

This result can be shown for any norm in $\mathbf{R}^{d(m+1)}$.

Proof Define

$$M_n(\phi)=\frac{1}{2N}\sum_{k=1}^{m}\|\mathbf{Y}_k-\mathbf{X}_k(\theta_0+\delta_k)\|_2^2+\lambda_n\sum_{k=1}^{m}\|\delta_k\|_1+c\lambda_n\|\theta_0\|_1$$

and

$$\widetilde{M}_n(\phi) = \frac{1}{2m}\sum_{k=1}^{m}(\theta_0+\delta_k-\theta_k^*)^T\Sigma_k(\theta_0+\delta_k-\theta_k^*) + \lambda_n\sum_{k=1}^{m}\|\delta_k\|_1 + c\lambda_n\|\theta_0\|_1,$$

where, $\phi = (\theta_0, \delta_1, \ldots, \delta_m)$. We define $K_M \subset \mathbf{R}^{d(m+1)}$ to be the closed ball around $\phi^* = (\theta_0^*, \theta_1^* - \theta_0^*, \ldots, \theta_m^* - \theta_0^*)$ with radius M. Since, for each $\phi \in K_M$, $M_n(\phi) \to \widetilde{M}_n(\phi)$ and $\sup_{\phi\in K_M}\widetilde{M}_n(\phi)$ is finite, $\sup_{\phi\in K_M}|M_n(\phi) - \widetilde{M}_n(\phi)| \to 0$ in probability.

For any $\epsilon > 0$,

$$\inf_{\|\phi-\phi^*\|_2\geq\epsilon}\widetilde{M}_n(\phi) > \widetilde{M}_n(\phi^*).$$

So, by argmin continuous mapping theorem

$$\left\|\arg\min_{\phi\in K_M} M_n(\phi) - \arg\min_{\phi\in K_M}\widetilde{M}_n(\phi)\right\|_2 \to 0$$

in probability. Taking the radius $M \to \infty$ we get

$$\left\|\arg\min_{\phi\in \mathbf{R}^{d(m+1)}} M_n(\phi) - \arg\min_{\phi\in \mathbf{R}^{d(m+1)}}\widetilde{M}_n(\phi)\right\|_2 \to 0$$

As $n \to \infty$, $\lambda_n \to 0+$. Hence, in the minimization problem of $\widetilde{M}_n(\phi)$ $\frac{1}{2m}\sum_{k=1}^{m}(\theta_0 + \delta_k - \theta_k^*)^T\Sigma_k(\theta_0+\delta_k-\theta_k^*)$ becomes more and more important, and in the limiting case this becomes primary objective. Looking at the loss for primary objective we see that optimum achieved at ϕ for which $\theta_0 + \delta_k = \theta_k^*$ for all $k \in [m]$. Hence, $\arg\min_{\phi\in\mathbf{R}^{d\times(m+1)}}\widetilde{M}_n(\phi) \to \arg\min_{\phi\in A}\sum_{k=1}^{m}\|\delta_k\|_1 + c\|\theta_0\|_1$, where, $A = \{\phi : \theta_0 + \delta_k = \theta_k^*\}$.

■

Appendix B. Supplementary Results

Lemma 12 *Under the Assumption 1 the followings hold.*

1. *The debiased lasso estimator $\widetilde{\theta}^d$ can be decomposed as*

$$\widetilde{\theta}^d = \theta^* - \widehat{\Theta}\nabla\ell(\theta^*) + \Delta$$

where, for and some $c_1 > 0$,

$$\mathbb{P}\left(\|\Delta\|_\infty > c_1 K s_0 \log d/n\right) \leq o(d^{-1}).$$

2. *For some $c_2 > 0$,*

$$\mathbb{P}\left(\textit{for some } j \in [d],\ \|\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}\|_1 > c_2\max\{Ks_j, K^2s_0\}\sqrt{\log d/n}\right) \leq o(d^{-1}).$$

3. *There is some $c_3 > 0$ such that $\|\Theta_{j,\cdot}\|_1 \leq c_3\sqrt{s_j}$ hold for any $j \in [d]$.*

Proof of Lemma 12. For the proof of 2. readers are suggested to see Theorem 3.2. in van de Geer et al. (2014).

Denote $s^* = \max\{s_1, \ldots, s_d\}$. To verify 1. we start with the Taylor expansion of $\widetilde{\theta}^d$:

$$\begin{aligned}
\widetilde{\theta}^d &= \widetilde{\theta} - \widehat{\Theta}\nabla\ell(\widetilde{\theta}) \\
&= \widetilde{\theta} - \widehat{\Theta}\frac{1}{n}\sum_{i=1}^{n}\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\widetilde{\theta}) \\
&= \widetilde{\theta} - \widehat{\Theta}\frac{1}{n}\sum_{i=1}^{n}\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta^*)\mathbf{x}_i - \widehat{\Theta}\frac{1}{n}\sum_{i=1}^{n}\ddot{\rho}(\mathbf{y}_i, \tilde{a}_i)\mathbf{x}_i\mathbf{x}_i^T(\widetilde{\theta} - \theta^*) \\
&= \theta^* - \widehat{\Theta}\nabla\ell(\theta^*) + (I - \widehat{\Theta}\mathbf{M})(\widetilde{\theta} - \theta^*)
\end{aligned}$$

where, $\tilde{a}_i$ is some number between $\mathbf{x}_i^T\widetilde{\theta}$ and $\mathbf{x}_i^T\theta^*$, and $\mathbf{M} = \frac{1}{n}\sum_{i=1}^{n}\ddot{\rho}(\mathbf{y}_i, \tilde{a}_i)\mathbf{x}_i\mathbf{x}_i^T$. We give a high probability ℓ_∞ bound for $\Delta = (I - \widehat{\Theta}\mathbf{M})(\widetilde{\theta} - \theta^*)$. By triangle inequality

$$\begin{aligned}
\|(I - \widehat{\Theta}\mathbf{M})(\widetilde{\theta} - \theta^*)\|_\infty &\leq \left\|\left(I - \widehat{\Theta}\nabla\ell(\widetilde{\theta})\right)(\widetilde{\theta} - \theta^*)\right\|_\infty + \left\|\widehat{\Theta}\left(\nabla^2\ell(\widetilde{\theta}) - \mathbf{M}\right)(\widetilde{\theta} - \theta^*)\right\|_\infty \\
&\leq \underbrace{\max_j \left\|e_j^T - \widehat{\Theta}_{j,\cdot}\nabla^2\ell(\widetilde{\theta})\right\|_\infty \|\widetilde{\theta} - \theta^*\|_1}_{I} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\Theta}\mathbf{x}_i\|_\infty|(\ddot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\widetilde{\theta}) - \ddot{\rho}(\mathbf{y}_i, \tilde{a}_i))\mathbf{x}_i^T(\widetilde{\theta} - \theta^*)|}_{II}
\end{aligned}$$

From KKT condition for nodewise lasso we get

$$-\frac{1}{n}\mathbf{X}_{\widetilde{\theta},-j}(\mathbf{X}_{\widetilde{\theta},j} - \mathbf{X}_{\widetilde{\theta},-j}\widehat{\gamma}_j) + \lambda_j\widehat{z}_j = 0,$$

where, $\|\widehat{z}_j\|_\infty \leq 1$. This implies

$$-\frac{1}{n}\widehat{\gamma}_j^T\mathbf{X}_{\widetilde{\theta},-j}\mathbf{X}_{\widetilde{\theta}}\widehat{\Theta}_{j,\cdot}^T\widehat{\tau}_j^2 + \lambda_j\|\widehat{\gamma}_j\|_1 = 0$$

Now

$$\begin{aligned}
\widehat{\tau}_j^2 &= \frac{1}{n}\|\mathbf{X}_{\widetilde{\theta},j} - \mathbf{X}_{\widetilde{\theta},-j}\widehat{\gamma}_j\|_2^2 + \lambda_j\|\widehat{\gamma}_j\|_1 \\
&= \frac{1}{n}\mathbf{X}_{\widetilde{\theta},j}^T\mathbf{X}_{\widetilde{\theta}}\widehat{\Theta}_{j,\cdot}^T\widehat{\tau}_j^2 - \frac{1}{n}\widehat{\gamma}_j^T\mathbf{X}_{\widetilde{\theta},-j}\mathbf{X}_{\widetilde{\theta}}\widehat{\Theta}_{j,\cdot}^T\widehat{\tau}_j^2 + \lambda_j\|\widehat{\gamma}_j\|_1 \\
&= \frac{1}{n}\mathbf{X}_{\widetilde{\theta},j}^T\mathbf{X}_{\widetilde{\theta}}\widehat{\Theta}_{j,\cdot}^T\widehat{\tau}_j^2.
\end{aligned}$$

Hence $\frac{1}{n}\mathbf{X}_{\widetilde{\theta},j}^T\mathbf{X}_{\widetilde{\theta}}\widehat{\mathbf{\Theta}}_{j,\cdot}^T = 1$. This implies

$$\|e_j - \nabla^2\ell(\widetilde{\theta})\widehat{\Theta}_{j,\cdot}^T\|_\infty = \left\|\frac{1}{n\widehat{\tau}_j^2}\mathbf{X}_{\widetilde{\theta},-j}(\mathbf{X}_{\widetilde{\theta},j} - \mathbf{X}_{\widetilde{\theta},-j}\widehat{\gamma}_j)\right\|_\infty \leq \frac{\lambda_j}{\widehat{\tau}_j^2} \lesssim \frac{1}{\widehat{\tau}_j^2}\left(\frac{\log d}{n}\right)^{\frac{1}{2}}.$$

By van de Geer et al. (2014), Theorem 3.2,

$$|\widehat{\tau}_j^2 - \tau_j^2| \lesssim \left(\frac{\max\{Ks^*, K^2 s_0\}\log d}{n}\right)^{\frac{1}{2}}$$

with probability at least $1 - o(d^{-1})$. Thus $\max_j \left\| e_j^T - \widehat{\Theta}_{j,\cdot}\nabla^2\ell(\widetilde{\theta})\right\|_\infty \lesssim_{\mathbb{P}} \left(\frac{\log d}{n}\right)^{\frac{1}{2}}$ and by 6. in the Assumption 1

$$\max_j \left\| e_j^T - \widehat{\Theta}_{j,\cdot}\nabla^2\ell(\widetilde{\theta})\right\|_\infty \|\widetilde{\theta} - \theta^*\|_1 \lesssim_{\mathbb{P}} \frac{s_0 \log d}{n},$$

where, $\lesssim_{\mathbb{P}}$ denotes $\lesssim$ with probability at least $1 - o(d^{-1})$.

We turn our attention to (II). We see that

$$\begin{aligned}\|\widehat{\Theta}\mathbf{X}^T\|_\infty &\leq \max_j \|\widehat{\Theta}_{j,\cdot}\mathbf{X}^T\|_\infty \lesssim \max_j \|\widehat{\Theta}_{j,\cdot}\mathbf{X}_{\theta^*}^T\|_\infty \\ &\leq \max_j \frac{1}{\widehat{\tau}_j^2}\|\mathbf{X}_{\theta^*,j} - \mathbf{X}_{\theta^*,-j}\widehat{\gamma}_j\|_\infty.\end{aligned}$$

Again by van de Geer et al. (2014), Theorem 3.2,

$$\begin{aligned}&\lesssim_{\mathbb{P}} \max_j \frac{1}{\tau_j^2}\|\mathbf{X}_{\theta^*,j} - \mathbf{X}_{\theta^*,-j}\widehat{\gamma}_j\|_\infty \\ &\leq \max_j \frac{1}{\tau_j^2}\|\mathbf{X}_{\theta^*,j} - \mathbf{X}_{\theta^*,-j}\gamma_j\|_\infty \\ &\quad + \frac{1}{\tau_j^2}\|\mathbf{X}_{\theta^*,-j}\|_\infty\|\widehat{\gamma}_j - \gamma_j\|_1,\end{aligned}$$

which, by 1. in the Assumption 1 and van de Geer et al. (2014), Theorem 3.2,

$$\lesssim_{\mathbb{P}} K + K \times \max\{Ks^*, K^2 s_0\}\sqrt{\frac{\log d}{n}}.$$

Now,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \|\widehat{\Theta}\mathbf{x}_i\|_\infty |(\ddot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\widetilde{\theta}) - \ddot{\rho}(\mathbf{y}_i, \tilde{a}_i))\mathbf{x}_i^T(\widetilde{\theta} - \theta^*)| \\ &\lesssim_{\mathbb{P}} K\frac{1}{n}\sum_{i=1}^n |(\ddot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\widetilde{\theta}) - \ddot{\rho}(\mathbf{y}_i, \tilde{a}_i))\mathbf{x}_i^T(\widetilde{\theta} - \theta^*)|\end{aligned}$$

which, by 5. and 6. in the Assumption 1, is at most

$$\lesssim \frac{K}{n}\|\mathbf{X}(\widetilde{\theta} - \theta^*)\|_2^2 \lesssim_{\mathbb{P}} \frac{Ks_0 \log d}{n}.$$

By union bound

$$\|\Delta\|_\infty \lesssim_{\mathbb{P}} \frac{(1+K)s_0 \log d}{n} = \mathcal{O}\big(Ks_0 \log d/n\big).$$

This shows assumption (A3).

From 3. in the Assumption 1 that the minimum eigen-value of Σ_θ^* is bounded away from zero for all d, we get that, for some $\kappa > 0$, which doesn't depend on d, the largest eigne-value of $\Theta^2 \leq \kappa$. Hence,

$$\begin{aligned}\|\Theta_{j,\cdot}\|_1^2 &\leq s_j \|\Theta_{j,\cdot}\|_1^2 \\ &= s_j e_j^T \Theta^2 e_j, \text{ where, } \{e_j\}_{j=1}^d \text{ is the standard basis of } \mathbf{R}^d \\ &\leq s_j \kappa.\end{aligned}$$

This shows 3. in lemma 12 ■

Proof of Lemma 2. We start with the linear decomposition, 1. in lemma 12, which give us

$$\|\widetilde{\theta}^d - \theta^*\|_\infty \leq \|\widehat{\Theta}\nabla\ell(\theta^*)\|_\infty + \|\Delta\|_\infty.$$

$$\begin{aligned}\|\widetilde{\theta}^d - \theta^*\|_\infty &\leq \|\widehat{\Theta}\nabla\ell(\theta^*)\|_\infty + \|\Delta\|_\infty \\ &\leq \underbrace{\|\Theta\nabla\ell(\theta^*)\|_\infty}_{I} + \underbrace{\|(\widehat{\Theta} - \Theta)\nabla\ell(\theta^*)\|_\infty}_{II} + \underbrace{\|\Delta\|_\infty}_{III}.\end{aligned}$$

From the fact that θ^* is the unique minimizer of $\mathbb{E}\rho(\mathbf{y}_i, \mathbf{x}_i^T\theta)$ we get

$$\mathbb{E}\left[\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta^*)\mathbf{x}_i\right] = 0.$$

Using 3. in lemma 12 and 7. in the Assumption 1 we have $\|\Theta\mathbf{x}_i\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i)\|_\infty \leq MK_3\sqrt{s^*}$ for some $M > 0$. Using Bernstein's inequality we get,

$$\mathbb{P}\left(\left|\sum_{i=1}^n \Theta_{j,\cdot}\mathbf{x}_i\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta^*)\right| > t\right) \leq 2\mathsf{exp}\left(-\frac{\frac{1}{2}t^2}{n\sigma_j^2 + \frac{1}{3}\sqrt{s^*}MK_3t}\right),$$

where, $\sigma_j^2 = \Theta_{j,\cdot}\mathbb{E}\left[\nabla\ell(\theta^*)\nabla\ell(\theta^*)^T\right]\Theta_{j,\cdot}^T$. For $t \leq \frac{3n\sigma_j^2}{\sqrt{s^*}MK_3}$ we get sub-Gaussian bound

$$\mathbb{P}\left(\left|\sum_{i=1}^n \Theta_{j,\cdot}\mathbf{x}_i\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta^*)\right| > t\right) \leq 2\mathsf{exp}\left(-\frac{t^2}{4n\sigma_j^2}\right).$$

By union bound

$$\mathbb{P}\left(\|\Theta\nabla\ell(\theta^*)\|_\infty > t\right) \leq 2d\mathsf{exp}\left(-\frac{nt^2}{4\sigma^2}\right),$$

where, $\sigma^2 = \max_j \sigma_j^2$. Setting $t = \sigma\sqrt{\frac{12\log d}{n}}$ we get $\mathbb{P}\left(\|\Theta\nabla\ell(\theta^*)\|_\infty > \sigma\sqrt{\frac{12\log d}{n}}\right) \leq \frac{2}{d^2}$. Since, for sufficiently large n we have $\sigma\sqrt{\frac{12\log d}{n}} \leq \frac{3n\sigma_j^2}{\sqrt{s^*}MK_3}$, such a choice for t is justified.

Since, $\left\{\mathbf{x}_i\dot{\rho}(\mathbf{y}_i, \mathbf{x}_i^T\theta^*)\right\}_{i=1}^n$ is bounded and zero mean random vectors, we get the following high probability bound for $\|\nabla\ell(\theta^*)\|_\infty$:

$$\text{For some } c_4 > 0,\ \mathbb{P}\left(\|\nabla\ell(\theta^*)\|_\infty > c_4\sqrt{\frac{\log d}{n}}\right) \leq \frac{1}{d^2}.$$

Let A the event that the followings hold:

1. $\|\Delta\|_\infty \geq c_1 K s_0 \log d/n$,
2. $\max_j \|\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}\|_1 \leq c_2 \max\{Ks^*, K^2 s_0\}\sqrt{\log d/n}$,
3. $\|\Theta \nabla \ell(\theta^*)\|_\infty \leq \sigma\sqrt{\frac{12\log d}{n}}$,
4. $\|\nabla \ell(\theta^*)\|_\infty \leq c_4\sqrt{\frac{\log d}{n}}$.

Then $\mathbb{P}(A) \geq 1 - o(d^{-1})$. Under the event A, $(I) \leq \sigma\sqrt{\frac{12\log d}{n}}$, and $(III) \leq c_1 K s_0 \log d/n$. We also notice that $(II) \leq \max_j \|\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}\|_1 \|\nabla \ell(\theta^*)\|_\infty \leq c_2 c_4 \max\{Ks^*, K^2 s_0\}\log d/n$.

Hence, $\mathbb{P}\left(\|\widetilde{\theta}^d - \theta^*\|_\infty \leq \sigma\sqrt{\frac{12\log d}{n}} + c\max\{Ks^*, K^2 s_0\}\frac{s^*\log d}{n}\right) \geq 1 - o(d^{-1})$.

■

Proof of Result 5. We shall prove uniqueness of the global parameter for each co-ordinates separately. Fix $j \in [d]$. For simplicity of the notation we ignore the index j, and denote $I_j, (\theta_k^*)_j, \mu_j$, and, η_j as I, θ_k^*, μ and, η, respectively.

For $x \in \mathbf{R}$ let us define $I_x = \{k : \theta_k^* \in [x - \eta, x + \eta]\}$, and $N_x = \sum_{k \in I_x} n_k$.

When $I_x = I$, we have

$$\sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - x\right) = (m - |I|)\eta^2 + \sum_{k \in I}(\theta_k^* - x)^2$$

which is uniquely minimized at $x = \mu$.

Under the case $I_x \neq I$ we must have $|\mu - x| > \delta$. We consider the case $\delta < |\mu - x| \leq 3\delta$. $|\mu - x| > \delta$ implies either x is larger or smaller than θ_k^*'s which are in I. We assume that $x > \theta_k^*$ for all $k \in I$. The other case will follow similarly. In that case,

$$(\theta_{\max}^* - \theta_k^*) \leq (2\delta) \wedge (x - \theta_k^*),$$

for all $k \in I$, where $\theta_{\max}^* = \max_{j \in I} \theta_k^*$. Since, $\eta > 2\delta$, we have

$$\sum_{k \in I} \Psi_\eta\left(\theta_k^* - x\right) \geq \sum_{k \in I} \Psi_\eta\left(\theta_k^* - \theta_{\max}^*\right).$$

Notice that, for $|x - \mu| \leq 3\delta$ we have $\Psi_\eta(\theta_k^* - x) = \eta^2$ whenever $k \notin I$. Hence,

$$\sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - x\right) \geq \sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - \theta_{\max}^*\right) > \sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - \mu\right),$$

where the last inequality follows form the fact that $I_x = I$ for $x = \theta_{\max}^*$.

Now, consider the case $|x - \mu| > 3\delta$, under which we have $I_x \cap I = \emptyset$.

Then

$$\sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - x\right) > 4|I|\delta^2$$

and

$$\sum_{k=1}^{m} \Psi_\eta\left(\theta_k^* - \mu\right) = (m - |I|)\eta^2 + \sum_{k \in I}(\theta_k^* - \mu)^2.$$

Since, the range of $\{\theta_k^*\}_{k\in I}$ is less than 2δ, we have $\sum_{k\in I}(\theta_k^*-\mu)^2 \le \delta^2|I|$. This implies

$$\begin{aligned}\sum_{k=1}^{m}\Psi_\eta\left(\theta_k^*-\mu\right) &\le (m-|I|)\eta^2+\delta^2|I| \\ &\le 4(m-|I|)\delta^2+|I|\delta^2 \\ &\le (4m-3|I|)\delta^2 \le 4|I|\delta^2\end{aligned}$$

and hence, μ is the unique minimizer in this case. ■

Proof of Lemma 6.

We start with remark 3 that under the assumption (ii) for sufficiently large n_k we have

$$\|\widetilde{\theta}_k^d-\theta_k^*\|_\infty \le 2\sigma\sqrt{\log d/n_k}$$

with probability at least $1-o(d^{-1})$. By union bound, with probability $1-o(1)$, simultaneously for all k we have

$$\|\widetilde{\theta}_k^d-\theta_k^*\|_\infty \le 2\sigma\sqrt{\log d/n_k}.$$

If $2\sigma\sqrt{\log d/n_k} \le \frac{1}{4}(\eta_j-2\delta)\wedge(\eta_j-\delta_2/2)$ for all k, then there exists $\hat{\delta}$ and $\hat{\delta}_2$ such that the following holds,

$$2\delta \le 2\hat{\delta} < \eta_j < \hat{\delta}_2/2 \le \delta_2/2.$$

and assumptions 4 holds with $\hat{\delta}$ and $\hat{\delta}_2$. Hence, by result 5 we have $(\widetilde{\theta}_0)_j = \frac{\sum_{k\in I_j}(\widetilde{\theta}_k^d)_j}{|I_j|}$.

Let $\gamma_k^{(j)} = \frac{1}{|I_j|}\mathbf{1}_{I_j}(k)$, where, $\mathbf{1}_A$ is the indicator function over the set A. From Lemma 12

$$\begin{aligned}(\widetilde{\theta}_0)_j &= \sum_{k=1}^{m}\gamma_k^{(j)}(\widetilde{\theta}_k^d)_j \\ &= \sum_{k=1}^{m}\gamma_k^{(j)}\left((\theta_k^*)_j-(\widehat{\Theta}_k)_{j\cdot}\nabla\ell_k(\theta_k^*)+(\Delta_k)_j\right) \\ &= (\theta_0^*)_j-\sum_{k=1}^{m}\gamma_k^{(j)}(\widehat{\Theta}_k)_{j\cdot}\nabla\ell_k(\theta_k^*)+\sum_{k=1}^{m}\gamma_k^{(j)}(\Delta_k)_j.\end{aligned}$$

Hence,

$$\begin{aligned}
\left|(\widetilde{\theta}_0)_j - (\theta_0^*)_j\right| &\le \left|\sum_{k=1}^m \gamma_k^{(j)} (\widehat{\Theta}_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right| + \left|\sum_{k=1}^m \gamma_k^{(j)} (\Delta_k)_j\right| \\
&\le \left|\sum_{k=1}^m \gamma_k^{(j)} (\Theta_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right| + \left|\sum_{k=1}^m \gamma_k^{(j)} \left((\widehat{\Theta}_k)_{j\cdot} - (\Theta_k)_{j\cdot}\right) \nabla \ell_k(\theta_k^*)\right| \\
&\quad + \left|\sum_{k=1}^m \gamma_k^{(j)} (\Delta_k)_j\right| \\
&\le \underbrace{\left|\sum_{k=1}^m \gamma_k^{(j)} (\Theta_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right|}_{I} \\
&\quad + \underbrace{\sum_{k=1}^m \gamma_k^{(j)} \left(\|(\widehat{\Theta}_k)_{j\cdot} - (\Theta_k)_{j\cdot}\|_1 \|\nabla \ell_k(\theta_k^*)\|_\infty + |(\Delta_k)_j|\right)}_{II}.
\end{aligned}$$

To bound (I) we notice that $\left\{(\Theta_k)_{j,\cdot} \mathbf{x}_{ki} \dot{\rho}(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T \theta_k^*)\right\}_{i\in[n_k], k\in[m]}$ are zero mean random variables bounded by $M\sqrt{s^*} > 0$. By Bernstein inequality,

$$\mathbb{P}\left(\left|\sum_{k=1}^m \gamma_k^{(j)} (\Theta_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right| > t\right) \le 2\mathsf{exp}\left(-\frac{\frac{1}{2}t^2}{\sum_{k=1}^m \left(\gamma_k^{(j)}\right)^2 \frac{\sigma_{jk}^2}{n_k} + \frac{1}{3}MtM\sqrt{s^*}}\right),$$

where $\sigma_{jk}^2 = \mathbb{E}\left[(\Theta_k)_{j,\cdot} \nabla \ell_k(\theta_k^*) \nabla \ell_k(\theta_k^*)^T (\Theta_k)_{j,\cdot}^T\right]$. Let $a_j^2 = \sum_{k=1}^m \left(\gamma_k^{(j)}\right)^2 \frac{\sigma_{jk}^2}{n_k} = \frac{1}{|I_j|^2} \sum_{k=1}^m \frac{\sigma_{jk}^2}{n_k}$. For $t \le \frac{3a_{jk}^2}{M\sqrt{s^*}}$ we get a probability bound with subgaussian tail

$$\mathbb{P}\left(\left|\sum_{k=1}^m \gamma_k^{(j)} (\Theta_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right| > t\right) \le 2\mathsf{exp}\left(-\frac{t^2}{4a_j^2}\right).$$

Letting $t = 2a_j\sqrt{3\log d}$ we get

$$\left|\sum_{k=1}^m \gamma_k^{(j)} (\Theta_k)_{j\cdot} \nabla \ell_k(\theta_k^*)\right| \le 2a_j\sqrt{3\log d}$$

with probability at least $1 - d^{-2}$. Taking union bound over all co-ordinates we get the above bound for each co-ordinates with probability at least $1 - d^{-1}$.

We shall apply Bennett's concentration inequality 15 on (I). Fix some $j \in [d]$. For $k \in [m],\ i \in [n_k]$ define $g_{ik} = \frac{\Theta_{j,\cdot} \nabla \rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T \theta_k^*)}{MK_3\sqrt{s^*}}$, and $a_{ki} = \frac{\gamma_k^{(j)} MK_3\sqrt{s^*}}{n_k}$. Then $\mathbb{E}g_{ki} = 0$ and $\mathbb{E}g_{ki}^2 \le \frac{\sigma^2}{M^2K_3^2 s^*} := \delta$ where, $\sigma^2 \ge \max_{k\in[m], j\in[d]} \Theta_{j,\cdot} \mathbb{E}[\nabla\rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta_k^*) \nabla\rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta_k^*)^T] \Theta_{j,\cdot}^T$, from assumption (A5). Also, $|g_{ki}| \le \frac{\|\Theta_{j,\cdot}\|_1 \|\nabla\rho(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta_k^*)\|_\infty}{MK_3\sqrt{s^*}} \le 1$. Let us define $a :=$

$(a_{ki}, k \in [m],\ i \in [n_k])$. Now we apply Bennett's inequality 15 on $\sum_{k=1}^m \gamma_k^{(j)}(\Theta_k)_{j\cdot}\nabla\ell_k(\theta_k^*) = \sum_{k=1}^m \sum_{i=1}^{n_k} a_{ki}g_{ki}$. We notice that for $t \leq \frac{\delta\|a\|_2^2}{\|a\|_\infty}e^2$ we have the following bound:

$$\mathbb{P}\left(\left|\sum_{k=1}^m\sum_{i=1}^{n_k} a_{ki}g_{ki}\right| > t\right) \leq 2\mathsf{exp}\left(-\frac{t^2}{2\delta\|a\|_2^2e^2}\right).$$

For $t = 2\sqrt{\delta\log d}\|a\|_2 e$ we get

$$\mathbb{P}\left(\left|\sum_{k=1}^m\sum_{i=1}^{n_k} a_{ki}g_{ki}\right| > 2\sqrt{\delta\log d}\|a\|_2 e\right) \leq \frac{2}{d^2}.$$

We need to confirm that such a choice of t is valid, *i.e.*, $2\sqrt{\delta\log d}\|a\|_2 \leq \frac{\delta\|a\|_2^2}{\|a\|_\infty}e^2$ or $\|a\|_\infty\sqrt{\log d} \leq c\sqrt{s^*}\|a\|_2$ for some c independent of m, d and $n_k, \in [m]$. We notice that

$$\|a\|_2 = \frac{MK_3\sqrt{s^*}}{|I_j|}\sqrt{\sum_{k\in I_j}\frac{1}{n_k}} \geq \frac{1}{\sqrt{|I_j|n_{\max}}}, \text{and } \|a\|_\infty \leq \frac{MK_3\sqrt{s^*}}{n_{\min}|I_j|}$$

and hence,

$$\frac{\sqrt{s^*}\|a\|_2}{\|a\|_\infty\sqrt{\log d}} \geq \sqrt{\frac{s^*n_{\min}^2|I_j|}{n_{\max}\log d}} \geq \sqrt{\frac{n_{\min}^2}{s^{*2}n_{\max}\log d}}.$$

From assumption (iii) we get that $\sqrt{\frac{n_{\min}^2}{s^{*2}n_{\max}\log d}}$ asymptotically goes to infinity. Hence, such a choice of t is valid.

We notice that $\|a\|_2$ is dependent of j. We define $b_j =: 2\sqrt{\delta}\|a\|_2 e \leq \frac{2e\sigma}{|I_j|}\sqrt{\sum_{k\in I_j}\frac{1}{n_k}}$.

To get a bound for (II) we see that $\max_j \|(\widehat{\Theta}_k)_{j\cdot} - (\Theta_k)_{j\cdot}\|_1 \lesssim \max\{Ks_k^*, K^2s_{k,0}\}\sqrt{\log d/n_k}$, and $\|\Delta_k\|_\infty \lesssim Ks_{k,0}\log d/n_k$. We also notice that $\{\mathbf{x}_{ki}\dot{\rho}(\mathbf{y}_{ki}, \mathbf{x}_{ki}^T\theta_k^*)\}_{i\in[n_k],k\in[m]}$ are mean zero bounded random vectors. Hence, by Bernstein inequality we can show that for sufficiently large n_k's we can get a bound

$$\|\nabla\ell_k(\theta_k^*)\|_\infty \lesssim \sqrt{\frac{\log d}{n_k}}$$

with probability at least $1 - d^{-2}$. Hence, by union bound we get

$$\|(\widehat{\Theta}_k)_{j\cdot} - (\Theta_k)_{j\cdot}\|_1\|\nabla\ell_k(\theta_k^*)\|_\infty + |(\Delta_k)_j| \lesssim \frac{\max\{Ks_{\max}^*, K^2s_{0,\max}\}\log d}{n_{\min}} \text{ for all } j, k$$

with probability at least $1 - o(1)$.

Again by union bound, we get

$$|(\widetilde{\theta}_0)_j - (\theta_0^*)_j| \leq b_j\sqrt{\log d} + C\frac{\max\{Ks_{\max}^*, K^2s_{0,\max}\}\log d}{n_{\min}}, \text{ for all } j \tag{B.1}$$

with probability at least $1 - o(1)$, where $C > 0$ is some constant. Since, $b_j \geq \sigma' \sqrt{\frac{1}{mn_{\max}}}$, for some σ', from assumption (iv) we get $\frac{\max\{Ks^*_{\max}, K^2 s_{0,\max}\} \log d}{n_{\min}} \lesssim b_j \sqrt{\log d}$. Hence, for sufficiently large $\frac{n^2_{\min}}{n_{\max}}$, the above high probability bound reduces to

$$|(\widetilde{\theta}_0)_j - (\theta^*_0)_j| \leq b_j \sqrt{\log d}$$

with probability at least $1 - o(1)$.

Under the assumption (iii) we have $b_j \leq 2e\sigma \sqrt{\frac{1}{|I_j|^2} \sum_{k \in I_j} \frac{1}{n_k}}$, which gives us the first result.

For the second inequality we get a bound for each $\widetilde{\theta}^d_k$ of the form:

$$\|\widetilde{\theta}^d_k - \theta^*_k\|_\infty \leq 2\sigma \sqrt{\frac{2 \log d}{n_k}} + C_k \frac{\max\{Ks^*_k, K^2 s_{k,0}\} \log d}{n_k}, \text{ for each } k$$

with probability at least $1 - o(1)$. Form the above bound and (B.1) we get

$$\begin{aligned} \|\widetilde{\delta}_k - \delta^*_k\|_\infty \leq\ & 2\sigma \sqrt{\frac{2 \log d}{n_k}} + C_k \frac{\max\{Ks^*_{\max}, K^2 s_{0,\max}\} \log d}{n_k} \\ & + 2e\sigma \sqrt{\frac{\log d}{|I_j|^2} \sum_{k \in I_j} \frac{1}{n_k}} + C \frac{\max\{Ks^*_k, K^2 s_{k,0}\} \log d}{n_{\min}}. \end{aligned}$$

hence, for sufficiently large n_k's, we have

$$\|\widetilde{\delta}_k - \delta^*_k\|_\infty \leq 4\sigma \sqrt{\frac{\log d}{n_k}} \text{ for all } k,$$

with probability at least $1 - o(1)$.

■

Result 13 *For each $k \in [m]$ let $\|\widetilde{\theta}_k - \theta^*_k\|_\infty \leq \xi$, where $\xi < \min_j \frac{1}{4}(\eta_j - 2\delta_j) \wedge (\delta_{2j}/2 - \eta_j)$. Then the assumptions 4 are satisfied for $\{\widetilde{\theta}_k\}_{k=1}^m$ for $\{\widetilde{\delta}_j = \delta_j + \xi/2\}_{j=1}^d$, $\{\widetilde{\delta}_{2j} = \delta_{2j} - 2\xi\}_{j=1}^d$ and $\{\eta_j\}_{j=1}^d$.*

Proof Let $j \in [d]$. Consider the same I_j as in assumption 4, (i). Let $\widetilde{\mu} = \frac{1}{I_j} \sum_{k \in I_j} \widetilde{\theta}_k$. Notice that for $\|\widetilde{\theta}_k - \theta^*_k\|_\infty \leq \xi$ we have $|\widetilde{\mu} - \mu| \leq \xi$. We further notice that $2(\delta_j + 2\xi) < \eta_j < \frac{\delta_{2j} - 2\xi}{2}$. This implies assumption 4, (ii) and (iii) are satisfied with $\widetilde{\delta}_j = \delta_j + 2\xi$ and $\widetilde{\delta}_{2j} = \delta_{2j} - 2\xi$. ■

Result 14 (Bernstein's inequality) *Let $\{\mathbf{x}_i\}_{i=1}^n$ be independent random variables with $\mathbb{E}\mathbf{x}_i = 0$, $\mathbb{E}\mathbf{x}_i^2 \leq \sigma^2$. Suppose for some $\lambda_0 > 0$ and $M > 0$ the following holds for all $i \in [n]$:*

$$\mathbb{E}\mathsf{exp}(\lambda_0 |\mathbf{x}_i|) \leq M.$$

Then for any $a \in \mathbf{R}^n$ and $t \leq \frac{M\sigma^2\|a\|_2^2\lambda_0}{\|a\|_\infty}$ the following holds:

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i \mathbf{x}_i\right| > t\right) \leq 2\mathsf{exp}\left(-\frac{t^2}{2\|a\|_2^2\sigma^2 M}\right).$$

Lemma 15 (Bennett's Inequality) *Let $X_1, X_2, \ldots X_n$ be independent random variables such that (1) $\mathbb{E}X_i = 0$, (2) $\mathbb{E}X_i^2 \leq \delta$, and (3) $|X_i| \leq 1$. Then for any $a_1, a_2, \ldots, a_n \in \mathbf{R}$ and for any $t > 0$*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq \begin{cases} 2\mathsf{exp}\left(-\frac{t^2}{2\delta\|a\|_2^2 e^2}\right), & \text{if } t \leq t^* \\ 2\mathsf{exp}\left(-\frac{t}{4\|a\|_\infty}\log\left(\frac{t\|a\|_\infty}{\delta\|a\|_2^2}\right)\right), & \text{if } t > t^* \end{cases}$$

where $t^* = \frac{\delta\|a\|_2^2}{\|a\|_\infty}e^2$.

Proof Without loss of generality assume $\|a\|_\infty = 1$. Let $\lambda > 0$. Set $Y = \sum_{i=1}^n a_i X_i$. Then

$$\mathbb{E}\mathsf{exp}(\lambda Y) = \prod_{i=1}^n \mathbb{E}\mathsf{exp}(\lambda a_i X_i).$$

For any $y \in \mathbf{R}$, $e^y \leq 1 + y + \frac{y^2}{2}e^{|y|}$.

Hence

$$\begin{aligned} \mathbb{E}\mathsf{exp}(\lambda a_i X_i) &\leq 1 + \mathbb{E}\frac{\lambda^2 a_i^2 X_i^2}{2} e^{|\lambda a_i X_i|} \\ &\leq 1 + \frac{\lambda^2 a_i^2 \delta}{2} e^{\lambda} \\ &\leq \mathsf{exp}\left(\frac{\lambda^2 a_i^2 \delta}{2} e^{\lambda}\right) \end{aligned}$$

Therefor $\mathbb{E}\mathsf{exp}(\lambda Y) \leq \mathsf{exp}\left(\frac{\lambda^2\|a\|_2^2\delta}{2}e^{\lambda}\right)$.

By Markov's inequality

$$\mathbb{P}(Y > t) \leq \mathsf{exp}\left(\frac{\lambda^2\|a\|_2^2\delta}{2}e^{\lambda} - \lambda t\right).$$

We have to minimize $\phi(\lambda) = \lambda t - \frac{\lambda^2\|a\|_2^2\delta}{2}e^{\lambda}$.

(1) Consider the case $\lambda \leq 2$. Then $\phi(\lambda) \geq \lambda t - \frac{\lambda^2\|a\|_2^2\delta}{2}e^2$. minimization with respect to λ gives us $\phi(\lambda) \geq \lambda t - \frac{\lambda^2\|a\|_2^2\delta}{2}e^2 \geq \frac{t^2}{2\|a\|_2^2\delta e^2}$. Since the minimum attains at $\lambda = \frac{t}{\|a\|_2^2\delta e^2}$, under the condition $\lambda \leq 2$ we have $t \leq \delta\|a\|_2^2 2e^2$. Hence, for $t \leq \delta\|a\|_2^2 2e^2$ we get the upper bound

$$\mathbb{P}(Y > t) \leq \mathsf{exp}\left(-\frac{t^2}{2\|a\|_2^2\delta e^2}\right).$$

(2) Consider the case of $\lambda > 1$. This implies $\lambda < e^{\lambda}$. Then $\phi(\lambda) \geq \lambda t - \frac{\lambda \|a\|_2^2 \delta}{2} e^{2\lambda} = \lambda \left(t - \frac{\|a\|_2^2 \delta}{2} e^{2\lambda} \right)$. Choose λ_2 such that $\frac{\|a\|_2^2 \delta}{2} e^{2\lambda_2} = \frac{t}{2}$. Then $\phi(\lambda) \geq \lambda_2 t/2 = \frac{t}{4} \log \left(\frac{t}{\|a\|_2^2 \delta} \right)$. Also from $\lambda_2 > 1$ we get $\frac{1}{2} \log \left(\frac{t}{\|a\|_2^2 \delta} \right) > 1$ which gives us $t \geq \|a\|_2^2 \delta e^2$. Hence, for $t \geq \|a\|_2^2 \delta e^2$ we get the upper bound

$$\mathbb{P}(Y > t) \leq \mathsf{exp} \left(-\frac{t}{4} \log \left(\frac{t}{\|a\|_2^2 \delta} \right) \right).$$

So, there is an overlap in the interval $(\delta \|a\|^2 e^2, 2\delta \|a\|_2^2 e^2)$ under which we get both the tails. We just choose $t^* = \delta \|a\|_2^2 e^2$. ■

Appendix C. A weighted data integration

In this section we consider a setup of weighted data integration where the weight $w = (w_1, \ldots, w_m)$ $(\sum_{k=1}^m w_k = 1)$ are applied in integration step the debiased lasso estimators:

$$\left(\widetilde{\theta}_0^w \right)_j = \arg\min_x \sum_{k=1}^m w_k \Psi_{\eta_j} \left(\left(\widetilde{\theta}_k^d \right)_j - x \right). \tag{C.1}$$

The corresponding global parameter $(\theta_0^{w,*})$ is identified via weighted re-descending loss:

$$(\theta_0^*)_j = \arg\min_{x \in \mathbf{R}} \sum_{k=1}^m w_k \Psi_{\eta_j}((\theta_k^*)_j - x), \tag{C.2}$$

and the unique identification of such global parameter is guaranteed by a similar set of structural conditions as in Assumption 4. We state the conditions below:

Assumption 16 *(P1) Let I_j be the set of indices for $(\theta_k^*)_j$'s which are considered as non-outliers. We assume $(\sum_{k \in I_j} w_k) \geq 4/7$.*

(P2) Let $\mu_j = \left(\sum_{k \in I_j} w_k (\theta_k^*)_j \right) / (\sum_{k \in I_j} w_k)$. Let δ be the smallest positive real number such that $(\theta_k^*)_j \in [\mu_j - \delta, \mu_j + \delta]$ for all $k \in I_j$. We assume that none of the $(\theta_k^*)_j$'s are in the intervals $[\mu_j - 5\delta, \mu_j - \delta)$ or $(\mu_j + \delta, \mu_j + 5\delta]$.

(P3) Let $\delta_2 = \min_{k_1 \in I_j, k_2 \notin I_j} |(\theta_{k_1}^*)_j - (\theta_{k_2}^*)_j|$. Clearly, $4\delta < \delta_2$. We choose η_j such that $2\delta < \eta_j < \delta_2/2$.

Result 17 *Under the conditions (P1)-(P3) in Assumption 16, the objective functions $\sum_{k=1}^m w_k \Psi_{\eta_j} ((\theta_k^*)_j - \theta)$ is uniquely minimized at $(\theta_0^{*,w})_j = \mu_j$, for all j.*

Proof This proof is exactly same as the proof of Result 5 if we assume that there are w_k proportitions of θ_k^*. ■

Theorem 18 *Let the followings hold:*

(i) For any j, $\{(\theta_k^*)_j\}_{k=1}^n$ satisfy the assumption 16.

(ii) The datasets $\{\mathcal{D}_k\}_{k=1}^m$ satisfy the Assumption 1 uniformly over k

(iii) Let $s_{k,0}$ *be sparsity for* θ_k^* *and* s_k^* *being the maximum sparsity for the rows of the inverse of* $\mathbb{E}[\ell_k(\theta_k^*)]$. *Define* $s_{0,\max} = \max_k s_{k,0}$, $s^*_{\max} = \max_k s_k^*$ *and* $s_{\text{eff}} = \max\{Ks^*_{\max}, K^2 s_{0,\max}\}$. *We assume* $m = o\left(\frac{n^2_{\min}}{s^2_{\text{eff}} n_{\max} \log d}\right)$, *where,* $n_{\max} = \max_{k\in[m]} n_k$, *and* $n_{\min} = \min_{k\in[m]} n_k$.

Then for sufficiently large n_k *we have the following bound for the co-ordinates of* $\widetilde{\theta}_0$ *and* ℓ_∞ *bound for* $\widetilde{\delta}_k$*:*

$$\begin{aligned} \left|(\widetilde{\theta}_0^w)_j - (\theta_0^{w,*})_j\right| &\le 4\sigma\sqrt{\frac{\log d}{(\sum_{k\in I_j} w_k)^2}\sum_{k\in I_j}\frac{w_k^2}{n_k}},\ \textit{for all } j, \\ \textit{and } \|\widetilde{\delta}_k^w - \delta_k^{w,*}\|_\infty &\le 4\sigma\sqrt{\frac{\log d}{n_k}},\ \textit{for all } k, \\ \widetilde{\delta}_k^w = \widetilde{\theta}_k^d - \widetilde{\theta}_0^w,&\quad \delta_k^{w,*} = \theta_k^* - \theta_0^{w,*} \end{aligned} \tag{C.3}$$

with probability at least $1 - o(1)$.

Proof The proof is similar to the proof of Lemma 6 if we replace $\gamma_k^{(j)}$ by $\frac{w_k}{\sum_{k\in I_j} w_k}\mathbf{1}_{I_j}(k)$. ■

Appendix D. Computational results: generalized linear model

In this section we perform synthetic experiments similar to Section 5 on binary classification setup. The covariates are $d = 100$ dimensional and again generated from AR(1) model with correlation $\rho = 0.9$ and variance 0.25 ($x = (x_1, \ldots . x_d)$, $x_1 \sim N(0, 0.25)$ and $x_j = \rho x_{j-1} + \sqrt{1-\rho^2}\epsilon_j$, $\epsilon_j \sim N(0, 0.25)$), and the response y_k for k-th dataset is generated as:

$$y_{k,i} \sim \text{Bernoulli}(\sigma(\mathbf{x}_{k,i}^\top \beta_k^*)), \quad \sigma(t) = \frac{e^t}{e^t + 1}\,.$$

We preform the debiased lasso on logistic regression to get the local estimates and keep rest of the setup exactly same as before.

In Figure 9 compare the ℓ_2-error of global parameter estimates with corresponding global parameters (note that they are different). The behaviors are similar to the linear models. For a fair comparison with the same baseline, we examine the prediction performance of the global parameters on new test data (not seen before), which one can find in Figure 10. The results for the logistic regression model have similar behavior as the linear regression model.

Appendix E. Supplementary details for cancer cell line study

In the following section we provide the detail about cancer cell line dataset.

E.1 Data source and preprocessing

The Cancer Cell Line Encyclopedia is a database on gene expression, genotype and drug sensitivity data. As covariates, we use the RNAseq TPM gene expression data DepMap

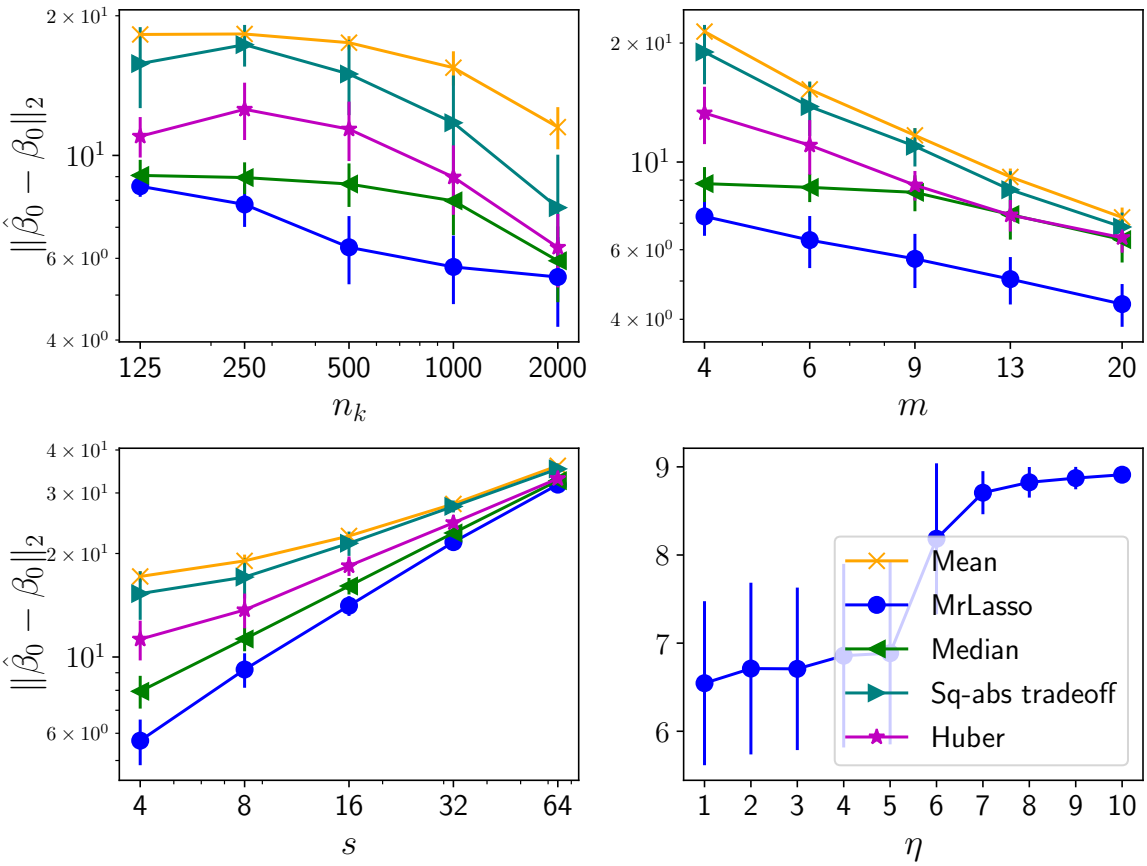

Figure 9: *Upper left*: Errorbar plots for ℓ_2-error in estimation of β_0 using different global estimators for varying sample sizes in each datasets (n_k) with $m = 5$ and $s = 5$. *Upper right*: For varying m with $n_k = 200$ and $s = 5$. *Lower left*: For different s with $n_k = 200$ and $m = 5$. *Lower right*: For different η with $n_k = 200$ and $s = 5$ and $m = 5$.

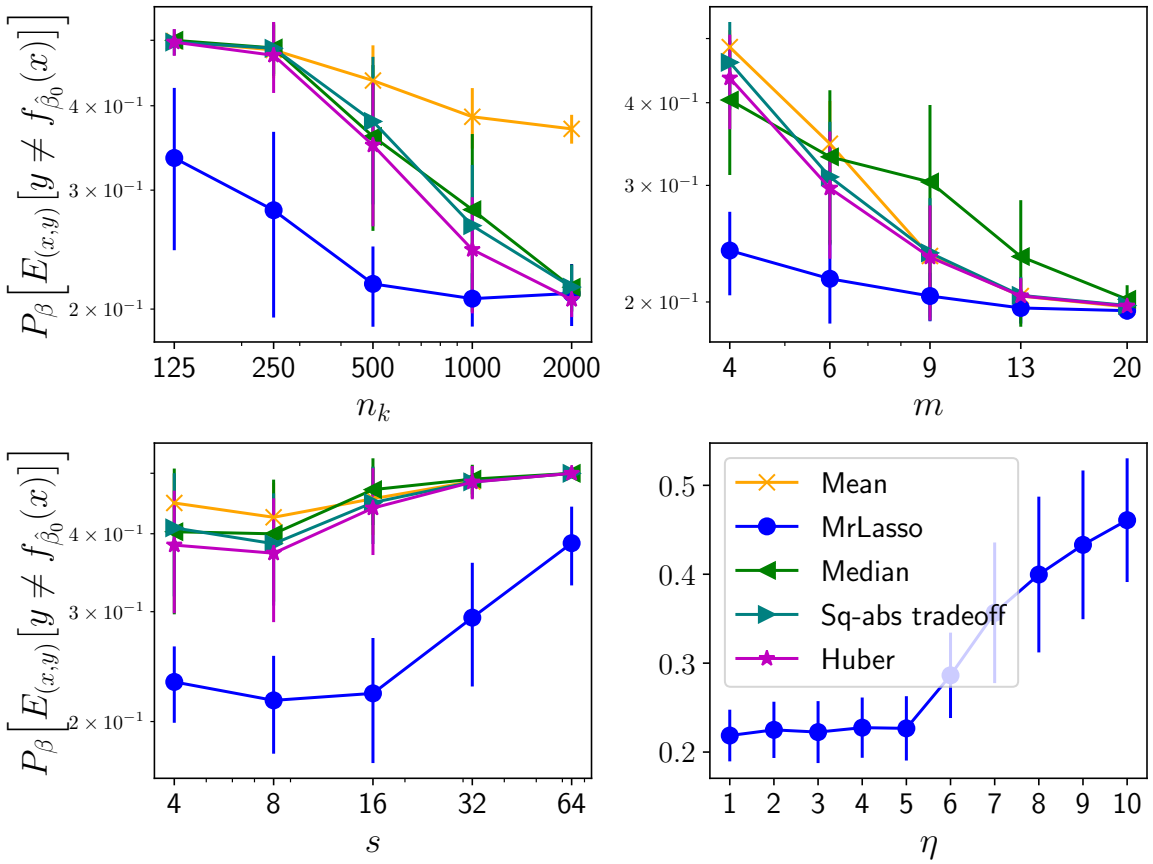

Figure 10: *Upper left*: Errorbar plots for prediction errors in test data using global estimators for varying sample sizes in each datasets (n_k) with $m = 5$ and $s = 5$. *Upper right*: For varying m with $n_k = 200$ and $s = 5$. *Lower left*: For different s with $n_k = 200$ and $m = 5$. *Lower right*: For different η with $n_k = 200$ and $s = 5$ and $m = 5$.

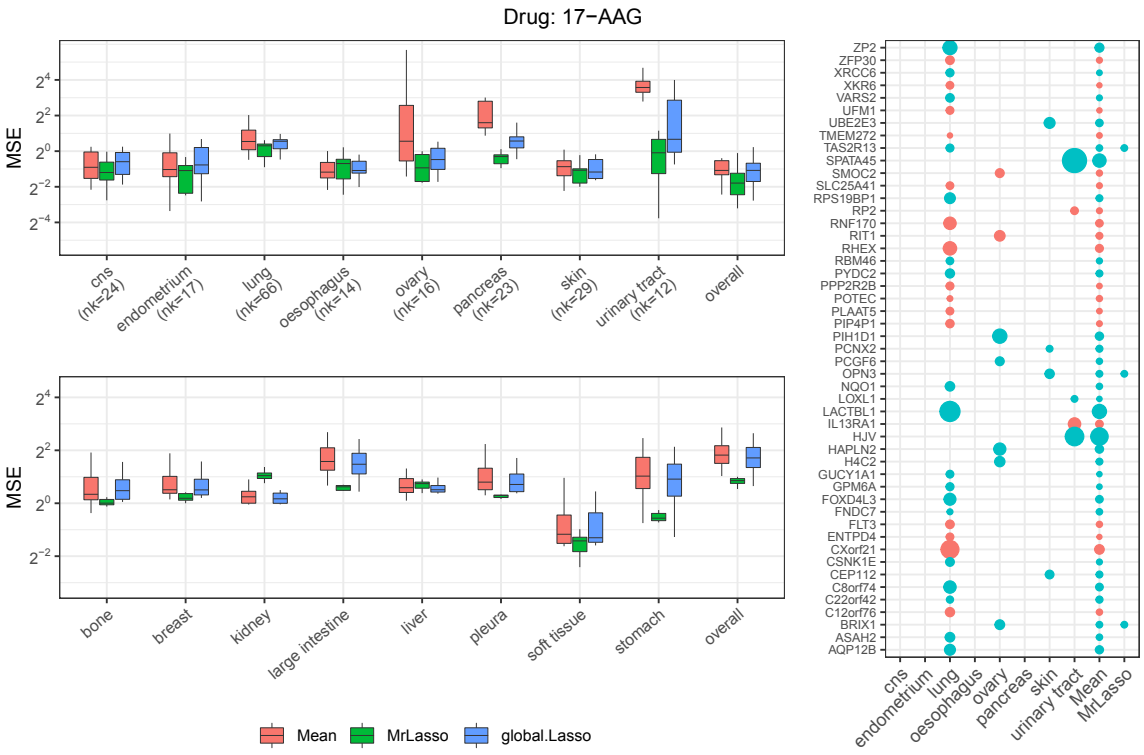

Figure 11: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug 17-AAG. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

(2021) for just protein coding genes (DepMap 21Q2 Public release) and the pharmacologic profiles for 24 anticancer drugs Consortium et al. (2015) across 504 across lines as the responses. The cell-lines in the gene expression data and the drug response data are matched using the pairs of cell line name and depmap id from the secondary screen dose response curve parameters file Corsello et al. (2019) (secondary-screen-dose-response-curve.csv). These data files are publicly available in DepMap portal[3]

The gene expression data has genetic expression for $d = 19177$ genes across 1379 cell lines. After matching them with the drug response data we only keep the cell lines for which the dose-response curve was fitted using sigmoid function. We use the area above dose-response curve (ActArea) as the drug response measure. Finally we end up with 482 cell lines in total. The codes are available in https://github.com/smaityumich/MrLasso.

E.2 Drug-response prediction plots

Next we present prediction error plots and gene selections for the drugs 17-AAG, Irinotecan, AZD0530, AEW541, ZD-6474, TKI258, RAF265 and PF2341066.

3. Depmap portal: https://depmap.org/portal/

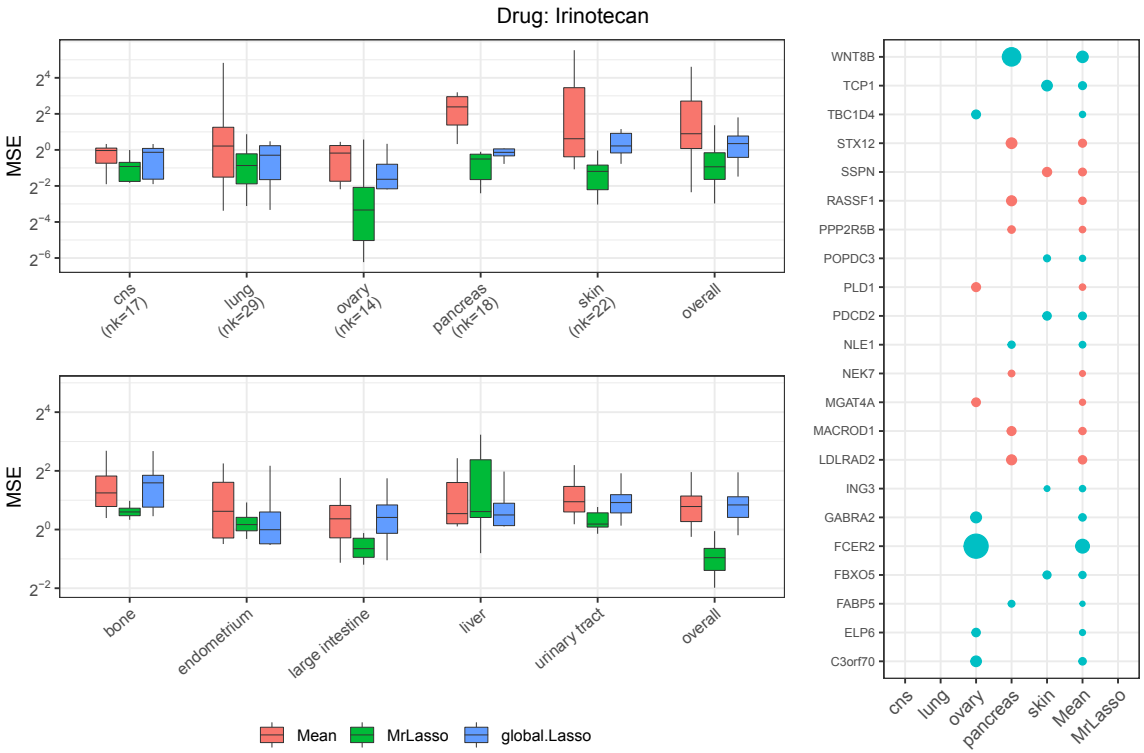

Figure 12: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug Irinotecan. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

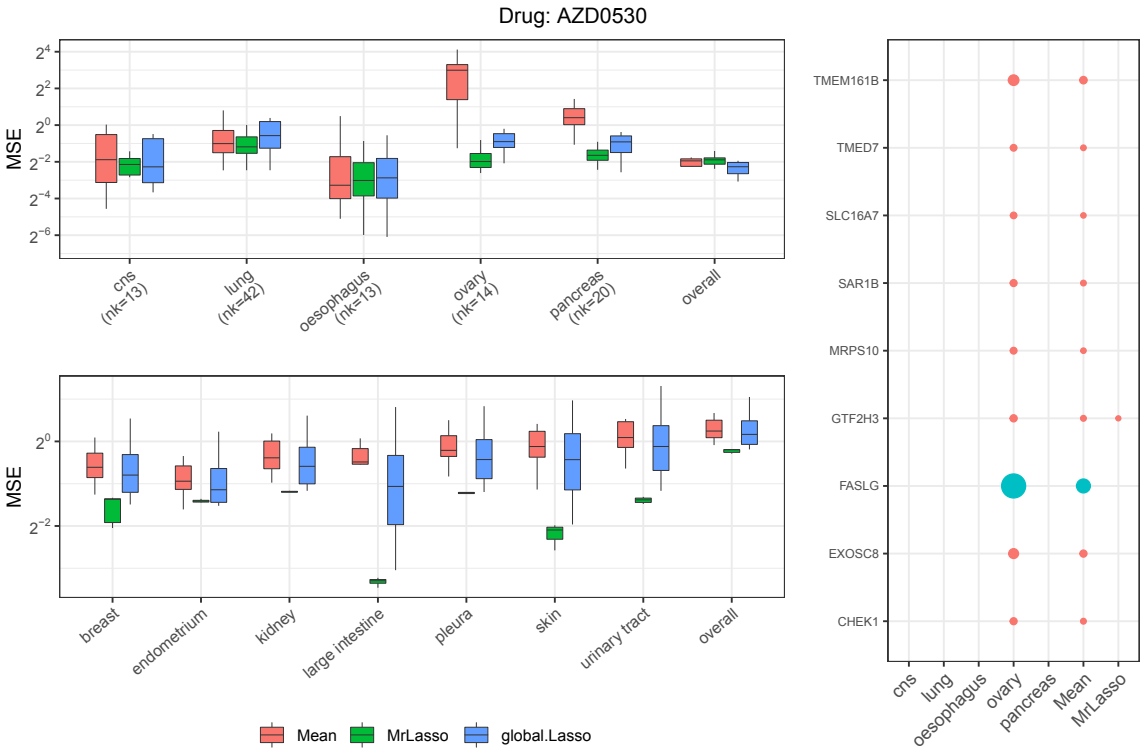

Figure 13: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug AZD0530. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

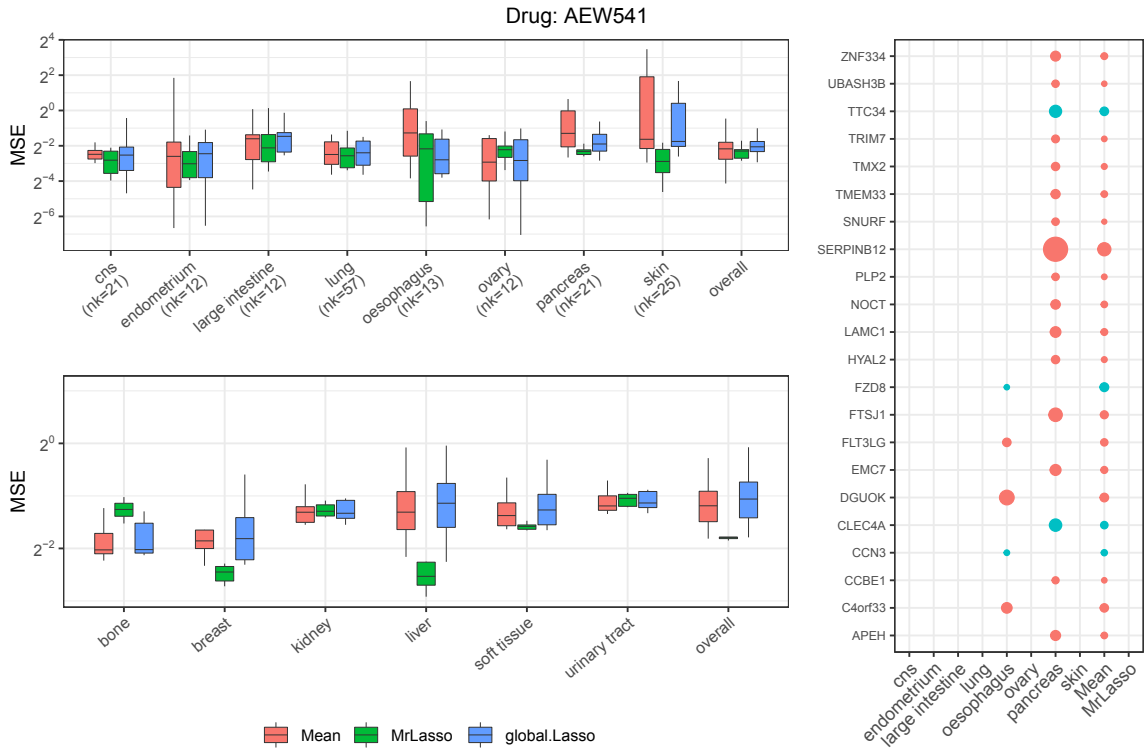

Figure 14: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug AEW541. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

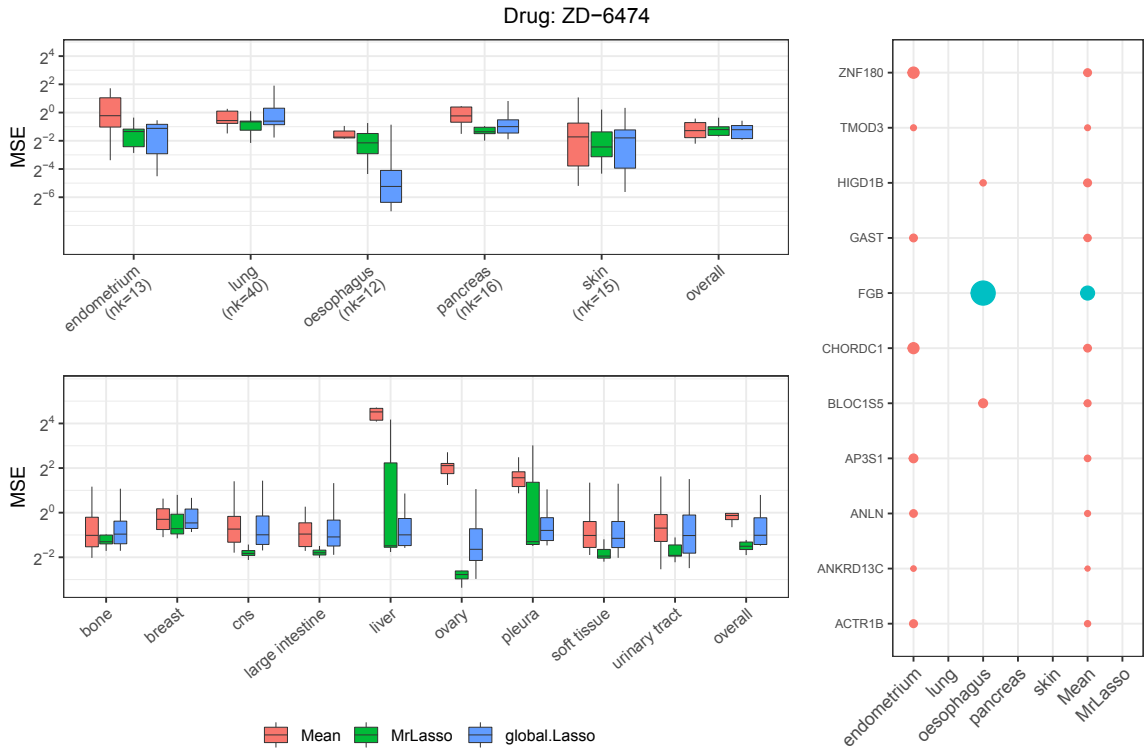

Figure 15: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug ZD-6474. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

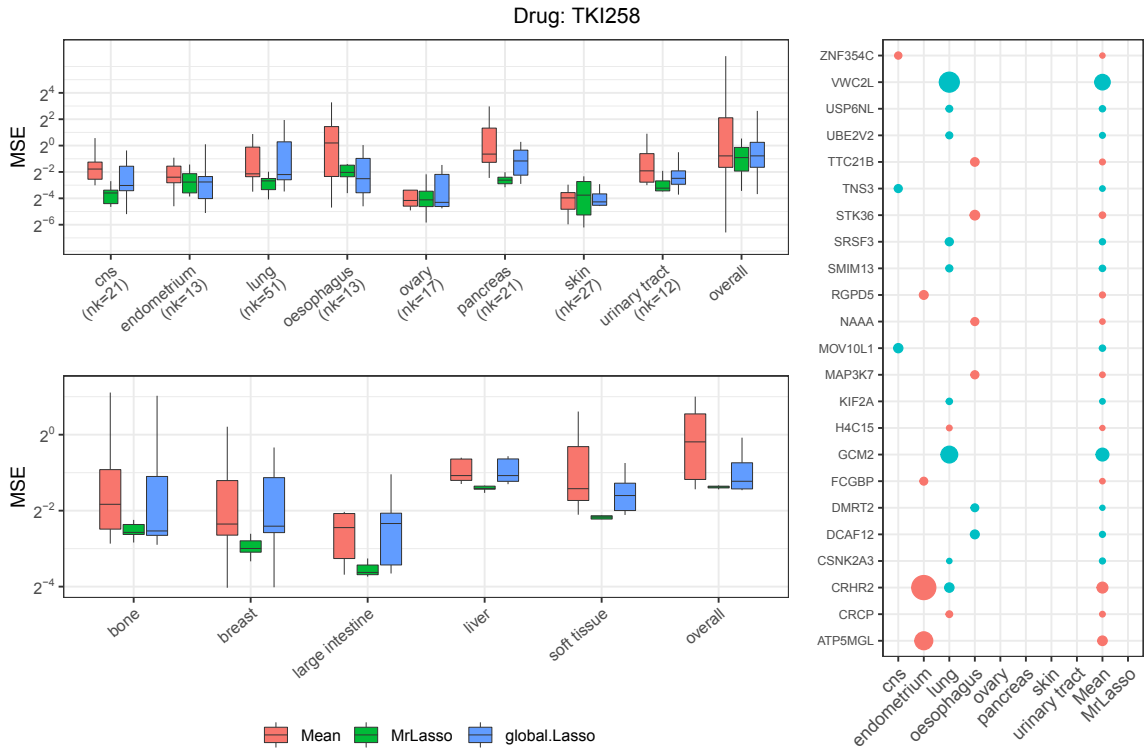

Figure 16: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug TKI258. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

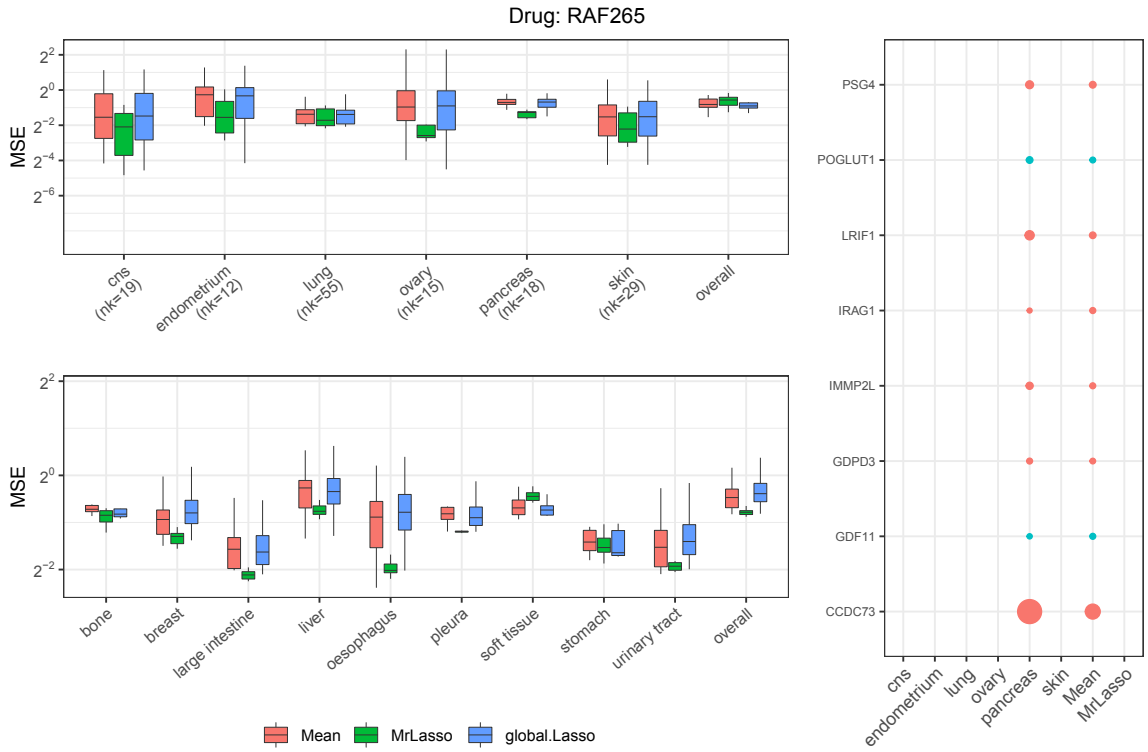

Figure 17: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug RAF265. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.

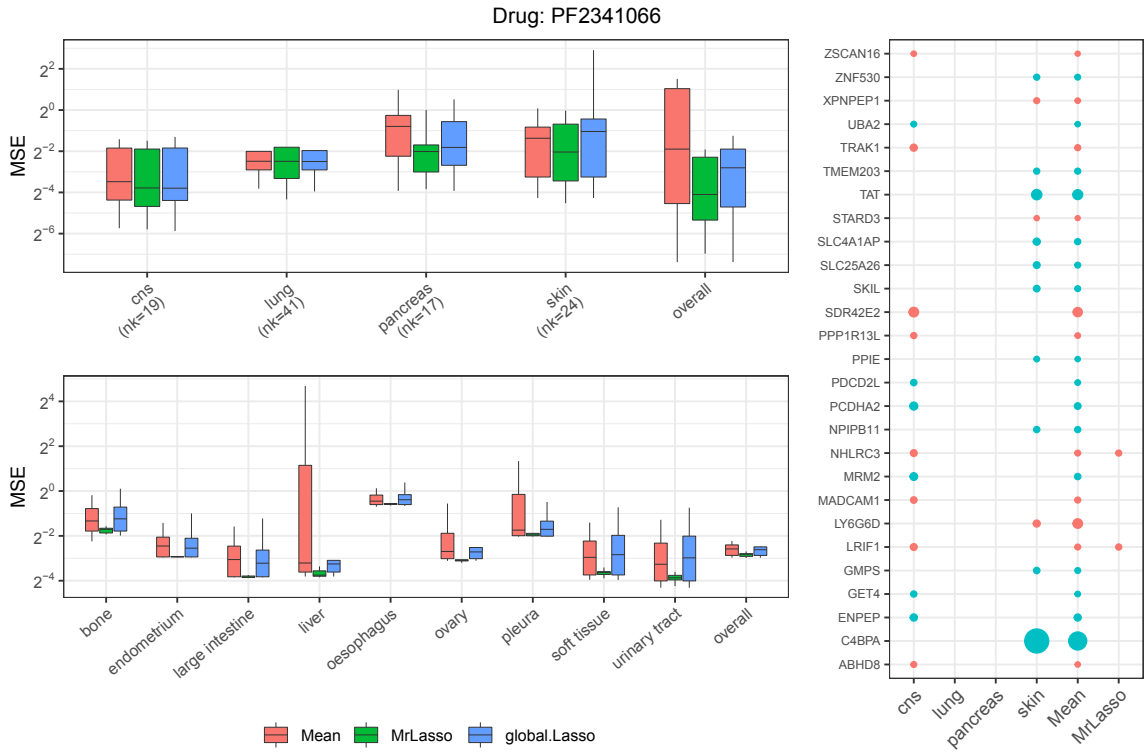

Figure 18: *Left:* Drug-response prediction mean squared errors (MSEs) for the most frequent cancer types (upper left) and rare cancer types (lower left panel) for the drug PF2341066. *Right:* Coefficient plot for the selected genes in lasso estimates for most frequent cancer types and ADELE and Mr Lasso.