Electronic Journal of Statistics
Vol. 18 (2024) 4499–4525
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2303

# Fitted value shrinkage

**Daeyoung Ham**

*School of Statistics, University of Minnesota,*
*e-mail:* ham00024@umn.edu

**and**

**Adam J. Rothman**

*School of Statistics, University of Minnesota,*
*e-mail:* arothman@umn.edu

**Abstract:** We propose a penalized least-squares method to fit the linear regression model with fitted values that are invariant to invertible linear transformations of the design matrix. This invariance is important, for example, when practitioners have categorical predictors and interactions. Our method has the same computational cost as ridge-penalized least squares, which lacks this invariance. We derive the expected squared distance between the vector of population fitted values and its shrinkage estimator as well as the tuning parameter value that minimizes this expectation. In addition to using cross validation, we construct two estimators of this optimal tuning parameter value and study their asymptotic properties. Our numerical experiments and data examples show that our method performs similarly to ridge-penalized least-squares.

**Keywords and phrases:** Invariance, penalized least squares, high-dimensional data.

Received October 2023.

## Contents

## 1. Introduction

We will introduce a new shrinkage strategy for fitting linear regression models, which assume that the measured response for $n$ subjects is a realization of the random vector

$$Y = X\beta + \varepsilon, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is the nonrandom known design matrix with ones in its first column and with values of the predictors in its remaining columns; $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients; and $\varepsilon$ has iid entries with mean zero and unknown variance $\sigma^2 \in (0, \infty)$.

We will consider fitting (1) in both low and high-dimensional settings, where the second scenario typically has $\text{rank}(X) < p$. If $\text{rank}(X) < p$, then it is well known that $\beta$ is not identifiable in (1), i.e. there exists a $\tilde{\beta} \neq \beta$ such that $X\beta = X\tilde{\beta}$. Similarly, if $\text{rank}(X) < p$, then there are infinitely many solutions to the least-squares problem: $\arg\min_{b\in\mathbb{R}^p} \|Y - Xb\|^2$. Given this issue (which is unavoidable in high dimensions), our inferential target is $X\beta$, which is the expected value of the response for the $n$ subjects.

To describe least squares estimators whether $\text{rank}(X) < p$ or $\text{rank}(X) = p$, we will use the reduced singular value decomposition of $X$. Let $q = \text{rank}(X)$. Then $X = UDV'$, where $U \in \mathbb{R}^{n\times q}$ with $U'U = I_q$; $V \in \mathbb{R}^{p\times q}$ with $V'V = I_q$; and $D \in \mathbb{R}^{q\times q}$ is diagonal with positive diagonal entries. The Moore–Penrose generalized inverse of $X$ is $X^- = VD^{-1}U'$ and a least-squares estimator of $\beta$ is $\hat{\beta} = X^-Y$. The vector of fitted values is $X\hat{\beta} = XX^-Y = P_XY$, where $P_X = XX^- = UU'$. If $\text{rank}(X) = p$, then $P_X = X(X'X)^{-1}X'$.

A nice property of this least-squares method is that its fitted values are invariant to invertible linear transformations of the design matrix. lose that we replace $X$ by $X_\bullet = XT$, where $T \in \mathbb{R}^{p\times p}$ is invertible. Then $X = X_\bullet T^{-1}$. So (1) is

$$Y = X\beta + \varepsilon = X_\bullet T^{-1}\beta + \varepsilon = X_\bullet\beta_\bullet + \varepsilon,$$

where $\beta_\bullet = T^{-1}\beta$. We estimate $X\beta = X_\bullet\beta_\bullet$ with $P_XY = P_{X_\bullet}Y$, so the fitted values did not change by changing $X$ to $X_\bullet$.

Fitting (1) by penalized least-squares has been studied by many scholars. Well-studied penalties include the ridge penalty [16], the bridge/lasso penalty [14, 26], the adaptive lasso penalty [34], the SCAD penalty [11], and the MCP penalty [31]. Unfortunately, these methods' fitted values are not invariant to invertible linear transformations of $X$. The lack of invariance is particularly problematic when a practitioner wants to find a set of linear combinations of a subset of the predictors to reduce multicollinearity, because a rotational change in the model matrix can substantially change the predicted values of the penalized methods that lack invariance. In addition, invariance to invertible linear transformations is also important in any application that includes categorical predictors (with three or more categories) and their interactions: the model fit should not change by changing the coding scheme used to represent these variables in the design matrix. The Group Lasso [29] and its variants with different standardization [7, 24] could also be used when there are categorical predictors, but these methods still lack invariance and our simulation and data examples illustrate their instability. This lack of invariance is also present in principal components regression [17], and partial least squares [27].

We illustrate that a change in the reference level coding of categorical predictors (see Section 5.2) and that a rotational transformation of the design matrix that removes zeros from the regression coefficient vector (see Section 5.3 and 5.5) leads the regression shrinkage methods that lack this invariance to perform worse. In contrast, our method performed the same before and after these transformations. We also propose a method to estimate the regression's error variance in high dimensions that may be of independent interest.

## 2. A new shrinkage method for linear regression with invariance

### *2.1. Method description*

To preserve the invariance to invertible linear transformations of the design matrix discussed in the previous section, we will use penalties that can be expressed as a function of the $n$-dimensional vector $Xb$, where $b$ is the optimization variable corresponding to $\beta$. To construct our estimator, we start with the following penalized least squares optimization:

$$\arg\min_{b\in\mathbb{R}^p} \left\{ \|Y - Xb\|^2 + \lambda \|Xb - \bar{Y}1_n\|^2 \right\}, \tag{2}$$

where $\bar{Y} = 1_n' Y/n$; $1_n' = (1, \ldots, 1) \in \mathbb{R}^n$; and $\lambda \in [0, \infty)$ is a tuning parameter. As $\lambda$ increases, the fitted values are shrunk towards the intercept-only model's fitted values $\bar{Y}1_n$. Let $\gamma = 1/(1+\lambda)$. We can express the optimization in (2) as

$$\hat{\beta}^{(\gamma)} \in \arg\min_{b\in\mathbb{R}^p} \left\{ \gamma \|Y - Xb\|^2 + (1-\gamma) \|Xb - \bar{Y}1_n\|^2 \right\}, \tag{3}$$

where $\gamma \in [0, 1]$. The $\in$ is used because there are infinitely many global minimizers for the optimization in (3) when $\text{rank}(X) < p$. To preserve invariance

and obtain the minimum 2-norm solution, we define our estimator as

$$\hat{\beta}^{(\gamma)} = X^{-}\left\{\gamma Y + (1-\gamma)\bar{Y}1_n\right\},$$

which is a global minimizer of (3) that uses the Moore-Penrose generalized inverse of $X$. Let $P_1 = 1_n(1_n{}'1_n)^{-1}1_n{}'$. The estimator of $X\beta$ is

$$X\hat{\beta}^{(\gamma)} = \gamma P_X Y + (1-\gamma)P_1 Y, \tag{4}$$

which is simply a convex combination of the least-squares fitted values $P_X Y$ and the intercept-only model's fitted values $P_1 Y = \bar{Y}1_n$. Since $P_X = P_{X_\bullet}$, where $X_\bullet = XT$ with $T \in \mathbb{R}^{p\times p}$ invertible, $X\hat{\beta}^{(\gamma)}$ is invariant to invertible linear transformations of $X$.

Due to its computational simplicity, $\hat{\beta}^{(\gamma)}$ is a natural competitor to ridge-penalized least squares, which lacks this invariance property. Both methods can be computed efficiently when $p$ is much larger than $n$ by using the reduced singular value decomposition of $X$ [15]. Specifically, they both cost $O(n\,\mathrm{rank}^2(X))$ floating-point operations. If $\gamma = 0$ for our method and $\lambda \to \infty$ for ridge-penalized least squares (without intercept penalization), then both procedures fit the intercept-only model.

We will derive an optimal value of $\gamma$ that minimizes $\mathbb{E}\|X\hat{\beta}^{(\gamma)} - X\beta\|^2$ and propose two estimators of it: one for low dimensions and one for high dimensions. We also explore using cross validation to select $\gamma$ when the response and predictor measurement pairs are drawn from a joint distribution. Conveniently, our results generalize to shrinkage towards a submodel's fitted values $P_{X_0}Y$, where $X_0$ is a matrix with a proper subset of the columns of $X$ (see Section C of the Supplementary material [13]).

### 2.2. Related work

Copas [9] proposed to predict a future value of the response for the $i$th subject with a convex combination of its fitted value (from ordinary least squares) and $\bar{Y}$. Although our methods are related, Copas [9] used a future-response-value prediction paradigm and did not establish a theoretical analysis of his approach.

Azriel and Schwartzman [2] study the estimation of $a'\beta$ when $p > n$ and $\beta$ is not sparse. They establish a condition for which $a'\beta$ is identifiable and show that using least-squares (with a pseudoinverse) has an optimal property when the errors are Gaussian. We prove that our shrinkage method outperforms least squares in same-$X$ prediction and we illustrate it in our numerical examples. Zhao et al. [32] also study the estimation of $a'\beta$, but they use procedures that are not invariant to invertible linear transformations of $X$.

## 3. Theoretical properties of the method

Given the lack of identifiability of $\beta$ in high dimensions, we investigate the estimation of the $n$-dimensional vector $X\beta$ with $X\hat{\beta}^{(\gamma)}$. This is an example of

same-X prediction [23]. It is related to predicting near $X$ when $p > n$ [8, Proposition 3.4]. In contrast to a random-design analysis, our fixed-design analysis allows a practitioner to control a subset of the columns of $X$, which would be essential for designed experiments.

Suppose that the linear regression model specified in (1) is true (this model did not specify an error distribution, just that they are iid mean 0 and variance $\sigma^2 \in (0, \infty)$). We define $\mu = X\beta$ and we assume that $\operatorname{rank}(X) \geq 2$ throughout the paper, so $X$ cannot correspond to an intercept-only model. Then we have the following result:

**Proposition 1.** *For all* $(n, p) \in \{1, 2, \ldots\} \times \{1, 2, \ldots\}$,

$$\mathbb{E}\|X\hat{\beta}^{(\gamma)} - X\beta\|^2 = \sigma^2(\gamma^2 r + 1 - \gamma^2) + (1-\gamma)^2\|\mu - P_1\mu\|^2.$$

The proof of Proposition 1 is in Section A.1 of the Supplementary material. When $\gamma = 1$, which is least squares, $\mathbb{E}\|X\hat{\beta}^{(1)} - X\beta\|^2 = \sigma^2\operatorname{rank}(X)$.

The right side of the equality in Proposition 1 is minimized when $\gamma = \gamma_{\text{opt}}$, where

$$\gamma_{\text{opt}} = \frac{\|\mu - P_1\mu\|^2}{\sigma^2(\operatorname{rank}(X) - 1) + \|\mu - P_1\mu\|^2}. \tag{5}$$

So the best our procedure could do is when $\|\mu - P_1\mu\|^2 = 0$ (that is, the intercept-only model is correct), in which case $\gamma_{\text{opt}} = 0$ and $\mathbb{E}\|X\hat{\beta}^{(0)} - X\beta\|^2 = \sigma^2$. The expression for $\gamma_{\text{opt}}$ enables us to construct a sample-based one-step estimator of it, which is more computationally efficiency than cross validation.

## 4. Selection of $\gamma$

### *4.1. Low-dimensional case*

Let $\hat{\sigma}^2 = \|Y - P_X Y\|^2/(n - \operatorname{rank}(X))$, which is an unbiased estimator of $\sigma^2$. To construct an estimator of $\gamma_{\text{opt}}$, we use the ratio of $\|P_X Y - P_1 Y\|^2 - \hat{\sigma}^2(\operatorname{rank}(X) - 1)$, which is an unbiased estimator of $\gamma_{\text{opt}}$'s numerator, to $\|P_X Y - P_1 Y\|^2$, which is an unbiased estimator of $\gamma_{\text{opt}}$'s denominator. This ratio estimator can be expressed as

$$\begin{aligned}\frac{\|P_X Y - P_1 Y\|^2 - \hat{\sigma}^2(\operatorname{rank}(X) - 1)}{\|P_X Y - P_1 Y\|^2} &= 1 - \frac{\hat{\sigma}^2(\operatorname{rank}(X) - 1)}{\|P_X Y - P_1 Y\|^2} \\ &= 1 - 1/F,\end{aligned}$$

where $F$ is the F statistic that compares the intercept-only model to the full model:

$$\begin{aligned}F &= \frac{(\|Y - P_1 Y\|^2 - \|Y - P_X Y\|^2)/(\operatorname{rank}(X) - 1)}{\|Y - P_X Y\|^2/(n - \operatorname{rank}(X))} \\ &= \frac{\|P_X Y - P_1 Y\|^2}{\hat{\sigma}^2(\operatorname{rank}(X) - 1)}.\end{aligned} \tag{6}$$

Since $F$ could be realized less than one (which corresponds to a fail-to-reject the intercept-only model situation), we define our estimator of $\gamma_{\text{opt}}$ to be

$$\hat{\gamma} = (1 - 1/F) \cdot 1(F > 1) \tag{7}$$

If the regression errors in (1) are Normal, $n > \text{rank}(X)$, and $\text{rank}(X) > 1$, then $F$ has a non-central F-distribution with degrees of freedom parameters $\text{rank}(X) - 1$ and $n - \text{rank}(X)$; and noncentrality parameter $\|\mu - P_1\mu\|^2/\sigma^2$. Larger realizations of $F$ correspond to worse intercept-only model fits compared to the full model, which makes $\hat{\gamma}$ closer to 1.

We also explore two additional estimators of $\gamma_{\text{opt}}$:

$$\hat{\gamma}_{90} = (1 - 1/F) \cdot 1(F \geq f_{0.9}), \tag{8}$$

$$\hat{\gamma}_{95} = (1 - 1/F) \cdot 1(F \geq f_{0.95}), \tag{9}$$

where $f_{90}$ and $f_{95}$ are the 0.9 and 0.95 quantiles of the central F-distribution with degrees of freedom $\text{rank}(X) - 1$ and $n - \text{rank}(X)$. These estimators may perform better when $\gamma_{\text{opt}}$ is near zero because they have a greater probability of estimating $\gamma_{\text{opt}}$ as zero than $\hat{\gamma}$ has.

Interestingly, Copas [9] proposed to predict a future response value for the $i$th subject with $(1 - \rho)\hat{\beta}' x_i + \rho\bar{Y}$, where $\rho \in [0, 1]$ is estimated and $\hat{\beta}$ is the ordinary least-squares estimator. They derived $1/F$ as an estimator of $\rho$ from the normal equations for the regression of $Y_{\text{new},i}$ on $\hat{\beta}' x_i$, $(i = 1, \ldots, n)$, where $Y_{\text{new},i}$ is an independent copy of $Y_i$. They also discussed using truncation to ensure their estimator of $\rho$ is in $[0, 1]$.

### *4.2. Consistency and the convergence rate of $\hat{\gamma}$*

We analyze the asymptotic performance of $\hat{\gamma}$ when the data are generated from (1) and $n$ and $p$ grow together. Define $r = \text{rank}(X)$ and $\delta^2 = \|\mu - P_1\mu\|^2$. The optimal tuning parameter value is a function of $r$ and $\delta^2$, so its value in the limit will depend on these sequences.

**Proposition 2.** *Assume that the data-generating model in (1) is correct, that the errors have a finite fourth moment, and that $r \geq 2$. If $p/n \to \tau \in [0, 1)$ and either $r \to \infty$ or $\delta^2 \to \infty$, then $\hat{\gamma} - \gamma_{\text{opt}} \to_P 0$ as $n \to \infty$.*

The proof of Proposition 2 is in Section A.2 of the Supplementary material. We see that consistency is possible whether the design matrix rank $r$ grows. If $r$ is bounded, then consistency requires $\delta^2 = \|\mu - P_1\mu\|^2 \to \infty$, which is reasonable even when the intercept-only model is a good approximation because $n$ is growing. One can also show consistency of $\hat{\gamma}_{90}$ and $\hat{\gamma}_{95}$ with $\delta^2 = o(r)$ added to the assumptions for Proposition 2.

Next, we establish a bound on the rate of convergence of $\hat{\gamma}$ with further assumptions on the design matrix $X$ and the error $\varepsilon$.

**Proposition 3.** *Suppose that the assumptions of Proposition 2 hold, that the errors in (1) are Gaussian, and that $r \geq 6$ is nondecreasing as $n \to \infty$. Then*

$$\hat{\gamma} - \gamma_{\text{opt}} = \begin{cases} O_P(r^{-1/2}) & \text{if } \delta^2 = O(r) \\ O_P((\delta^2/r)^{-3/4}) + O_P(n^{-1/2}(\delta^2/r)^{-1/2}) & \text{if } r \to \infty \text{ and } r = o(\delta^2) \\ O_P((\delta^2)^{-3/4}) + O_P(n^{-1/2}(\delta^2)^{-1/2}) & \text{if } r = O(1). \end{cases} \tag{10}$$

The proof of Proposition 3 is in Section A.3 of the Supplementary material. From the definition of $\gamma_{\text{opt}}$, we know that $\gamma_{\text{opt}} \to 0$ when $\delta^2 = o(r)$, in which case $\hat{\gamma} - \gamma_{\text{opt}} = O_P(n^{-1/2})$ provided that $r \asymp n$. On the other hand, $\gamma_{\text{opt}} \to 1$ when $r = o(\delta^2)$. For example, $\hat{\gamma} - \gamma_{\text{opt}} = O_P(n^{-1/2})$ provided that $\delta^2 \asymp n^{2/3}$ and $r$ is bounded. When $\delta^2 = o(r)$, $\hat{\gamma}_{90}, \hat{\gamma}_{95}$ and $\hat{\gamma}$ all have the same convergence rate bound.

### *4.3. Tuning parameter selection in high dimensions*

Estimating the unknown parameters in $\gamma_{\text{opt}}$ is challenging when $p > n$ and $\text{rank}(X) = n$. For example, it is impossible to estimate the regression's error variance $\sigma^2$ without assuming something extra about $\mu = X\beta$. This is because the data-generating model in (1) reduces to

$$Y = \mu + \varepsilon,$$

where $\mu$ has $n$ unknown free parameters and $\varepsilon$ has iid entries with mean zero and variance $\sigma^2$. So we have a sample size of 1 to estimate each $\mu_i$, which is not enough if we also want to estimate $\sigma^2$. An anonymous referee mentioned that if there are replicated rows in $X$ (which implies $\text{rank}(X) < n$), then one could estimate $\sigma^2$ using the measured responses corresponding to these the repeated rows.

We explore using cross-validation to choose a value of $\gamma$ that minimizes the total validation squared error in our numerical experiments. This cross-validation procedure implicitly assumes that the response and predictor measurement pairs for each subject are drawn from a joint distribution. As an alternative, we derive a high-dimensional estimator of $\gamma_{\text{opt}}$ that estimates $\sigma^2$ with an assumption about $\mu$. The following paragraphs introduce this estimator, which is not invariant to invertible linear transformations of $X$.

Recall that $\gamma_{\text{opt}} = \delta^2/(\sigma^2(\text{rank}(X) - 1) + \delta^2)$, where $\delta^2 = \|\mu - P_1\mu\|^2$. Since $P_X$ is an identity operator when $\text{rank}(X) = n$,

$$\mathbb{E}\|Y - P_1 Y\|^2 = \mathbb{E}\|P_X Y - P_1 Y\|^2 = \sigma^2(\text{rank}(X) - 1) + \delta^2.$$

So given an estimator $\check{\sigma}^2$ of $\sigma^2$, we study the following plug-in estimator of $\gamma_{\text{opt}}$:

$$\bar{\gamma}(\check{\sigma}^2) = \max\left(0, \frac{\|Y - P_1 Y\|^2 - \check{\sigma}^2(\text{rank}(X) - 1)}{\|Y - P_1 Y\|^2}\right), \tag{11}$$

where the truncation at 0 is necessary to ensure that $\bar{\gamma} \in [0, 1]$. We continue by describing the estimator of $\sigma^2$ that we will use in (11).

If one ignores invariance and assumes Gaussian errors, then one could simultaneously estimate $\beta$ and $\sigma^2$ by penalized likelihood with the same penalty used in (2). However, this joint optimization is not convex. We avoid this nonconvexity by modifying a reparametrized penalized Gaussian likelihood optimization problem proposed by Zhu [33]. Let $\eta = \sigma^{-1}$ and $\beta^* = \beta\eta$. We estimate these parameters with

$$(\hat{\beta}^*, \hat{\eta}) = \underset{(\beta^*,\eta)\in\mathbb{R}^p\times(0,\infty)}{\arg\min} \left\{ \frac{1}{2n}\|Y\eta - X\beta^*\|^2 - \log(\eta) + \alpha\|\beta^*_{-1}\|^2 \right\}, \tag{12}$$

where $\beta^* = (\beta^*_1, \ldots, \beta^*_p) = (\beta^*_1, \beta^*_{-1})$; and $\alpha \geq 0$ be the tuning parameter for the Ridge penalty. This choice of $\alpha$ was motivated by Liu et al. [22], who verified that ridge regression (with tuning parameter $\alpha$) can be used to consistently estimate $\sigma^2$ provided that $\alpha\|\beta\|^2 = o(1)$. However, Liu et al. [22] use a different estimator of $\sigma^2$ than the transformed solution to (12). We also examine other choices for $\alpha$ in the simulations (see section 5.4).

The reparametrized optimization problem in (12) is strongly convex with the following global minimizer:

$$\begin{aligned} \hat{\eta} &= \left(n^{-1}Y'(I - X(X'X + 2n\alpha M)^{-1}X')Y\right)^{-1/2}, \\ \hat{\beta}^* &= \hat{\eta}(X'X + 2n\alpha M)^{-1}X'Y, \end{aligned} \tag{13}$$

where $M = \mathrm{diag}(0, 1, 1, \ldots, 1) \in \mathbb{R}^{p\times p}$. Since $\eta = \sigma^{-1}$, which is estimated using (13), the corresponding estimator of $\sigma^2$ is

$$\check{\sigma}^2 = n^{-1}Y'(I - K)Y, \tag{14}$$

where $K = X(X'X + 2n\alpha M)^{-1}X'$. Using $\check{\sigma}^2$ in (11), we propose a high-dimensional estimator of $\gamma_{\mathrm{pot}}$ which is given by

$$\bar{\gamma} = \frac{Y'(I - P_1 - \frac{\mathrm{rank}(X)-1}{\mathrm{rank}(X)}(I - K))Y}{Y'(I - P_1)Y}, \tag{15}$$

where truncation at 0 is not required since the quadratic term on the numerator of (15) is non-negative definite. Liu et al. [22] proposed a bias corrected version of $\check{\sigma}^2$, since the uncorrected estimator does not converge to $\sigma^2$ under the assumptions for Theorem 1 of Liu et al. [22]. Their corrected estimator is $\check{\sigma}^2_c = C^{-1}\check{\sigma}^2$, where $C = 1 - \mathrm{tr}(K)/\mathrm{rank}(X)$. Using $\check{\sigma}^2_c$ in (11), we also propose alternative high-dimensional estimator of $\gamma_{\mathrm{opt}}$ by

$$\bar{\gamma}_c = \max\left(0, \frac{Y'(I - P_1 - \frac{\mathrm{rank}(X)-1}{\mathrm{rank}(X)-\mathrm{tr}(K)}(I - K))Y}{Y'(I - P_1)Y}\right). \tag{16}$$

First, we have the following consistency result for the corrected estimator $\bar{\gamma}_c$.

**Proposition 4.** *Assume that the data-generating model in (1) is correct, that error distribution has a finite fourth moment,* $\delta^2 = o(n)$*, and* $d_2(n\alpha)^{-1} = o_P(1)$*, where* $d_2$ *is the second-largest eigenvalue of* $XX'$*. Then* $\bar{\gamma}_c - \gamma_{\text{opt}} \to_P 0$ *as* $n \to \infty$*.*

The proof is in Section A.4 of the Supplementary material. We see that $\bar{\gamma}_c$ converges to $\gamma_{\text{opt}}$ when $\gamma_{\text{opt}} \to 0$. The assumption that $d_2(n\alpha)^{-1} = o_P(1)$ is met when $\alpha = n^{3/2}\left(2\|Y - P_1 Y\|^2\right)^{-1}$, $\delta^2 = o(n)$, and $d_2 = O(n)$ because

$$\begin{aligned} d_2(n\alpha)^{-1} &= 2n^{-5/2} d_2 \|Y - P_1 Y\|^2 \\ &= 2n^{-5/2} d_2 (\delta^2 + 2(\mu - P_1\mu)'\varepsilon + \varepsilon'(I - P_1)\varepsilon) \to_P 0, \end{aligned}$$

since $\|\varepsilon\|^2 = O_P(n)$.

We also established the following result for the uncorrected estimator $\bar{\gamma}$.

**Proposition 5.** *Assume that the data-generating model in (1) is correct, that the error distribution has a finite fourth moment, and at least one of the followings holds:*

- $\delta^2 = o(n)$ *and* $d_2(n\alpha)^{-1} = o_P(1)$*,*
- $n = o(\delta^2)$ *and* $n\alpha(d_n)^{-1} = o_P(1)$*,*

*where* $d_2$ *(resp.* $d_n$*) is the secondly largest (smallest) eigenvalue of* $XX'$*. Then* $\bar{\gamma} - \gamma_{\text{opt}} \to_P 0$ *as* $n \to \infty$*.*

The proof is in Section A.5 of the Supplementary material. Using the uncorrected error variance estimator $\breve{\sigma}^2$ enables us to prove consistency of $\bar{\gamma}$ when either $\gamma_{\text{opt}} \to 0$ or $\gamma_{\text{opt}} \to 1$. When, $\gamma_{\text{opt}} \to 0$, we have consistency provided that $d_2 = O(n)$ and $\alpha = n^t \left(2\|Y - P_1 Y\|^2\right)^{-1}$ with any $t \in (1, \infty)$. We test $t = 2, 3$ in Section 5.4. On the other hand, when $\gamma_{\text{opt}} \to 1$, we have consistency provided that $d_n \asymp n$ and $\alpha = n\left(2\|Y - P_1 Y\|^2\right)^{-1}$.

**Remark 1.** *Azriel [1] proved that "general" consistent estimation of the error variance* $\sigma^2$ *in high-dimensional linear regression is impossible without further assumptions in fixed-X settings. Their proof is based on a Bayes risk approach that exploits a prior that requires* $\delta^2 \sim n$*, which is not what we assume in Proposition 5 (*$\delta^2 = o(n)$ *or* $n = o(\delta^2)$*).*

## 5. Simulation studies

### *5.1. Low-dimensional experiments*

We conducted a lower-dimensional simulation study in which the data were generated from the linear regression subjects model (1) with $n = 300$ and $p \in \{75, 150, 250\}$. Also, $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, 1)$. The design matrix $X$ has ones in its first column and independent draws from $N_{p-1}(0, \Sigma)$ in the remaining

entries on each row, where $\Sigma_{jk} = 0.5^{|j-k|}$. We randomly generated the regression coefficient vector with the following equation:

$$\beta = X^{-}(1_p + \tau Z),$$

where $\tau \in \{0, 10^{-4}, 10^{-2}, 10^{-1}, 10^{-0.5}, 1, 10^{0.5}, 10^{1}, 10^{1.5}, 10^{2}\}$; and $Z$ is $N_p(0, I)$. Then

$$\mu - P_1\mu = (P_X - P_1)(1_p + \tau Z) = \tau(P_X - P_1)Z.$$

So $\tau$ controls the size of $\delta^2 = \|\mu - P_1\mu\|^2$.

We used 50 independent replications in each setting. In each replication, we measured the performance of each estimator $\hat{\beta}_{\text{est}}$ using the same-X loss:

$$n^{-1}\|X\beta - X\hat{\beta}_{\text{est}}\|^2. \tag{17}$$

The candidate estimators $\hat{\beta}_{\text{est}}$ that we considered were the following:

- **2n-G**: $L_2$-squared penalty with 10-fold cross validation for $\gamma$ (3).
- **2n-Or**: $L_2$-squared penalty using the oracle $\gamma_{\text{opt}}$ in (5).
- **2n-Es**: $L_2$-squared penalty using $\hat{\gamma}$ in (7).
- **2n-Es90**: $L_2$-squared penalty using $\hat{\gamma}_{90}$ in (8).
- **2n-Es95**: $L_2$-squared penalty using $\hat{\gamma}_{95}$ in (9).
- **2n-Rep**: $L_2$-squared penalty using $\check{\sigma}^2$ in (14), $\alpha = n(2\|Y - P_1Y\|^2)^{-1}$ in (12), and the corresponding $\bar{\gamma}_{\text{low}} = \max(0, 1-1/F_{\text{rep}})$, where $F_{\text{rep}} = \|P_XY - P_1Y\|^2/(\check{\sigma}^2(\text{rank}(X)-1))$.
- **O**: Ordinary least square (OLS) estimator given by

$$\hat{\beta}^{(1)} = X^{-}Y. \tag{18}$$

- **R**: Ridge-penalized least squares [16]

$$\hat{\beta}_{\text{Ridge}} = \text{argmin}_{b\in\mathbb{R}^p}\|Y - Xb\|^2 + \lambda\|b_{-1}\|^2, \tag{19}$$

  where $b = (b_1, b_{-1})$ with $b_{-1} = \mathbb{R}^{p-1}$; 10-fold cross validation for the selection of $\lambda$.
- **L**: Lasso-penalized least squares [26]

$$\hat{\beta}_{\text{LASSO}} = \text{argmin}_{b\in\mathbb{R}^p}\|Y - Xb\|^2 + \lambda\|b_{-1}\|_1, \tag{20}$$

  where 10-fold cross validation is used for the selection of $\lambda$.

For the methods that require cross validation, $\lambda$ and $\gamma$ were selected from $\{10^{-7+0.25j} : j = 0, 1, \cdots, 44\}$ and $\{\frac{k}{99} : k = 0, 1, \cdots, 99\}$, respectively. To facilitate the fairest comparison between our invariant methods and the ridge/lasso methods, we used the following standardization process, which is the default

process used by the R package `glmnet`: the ridge/lasso shrunken coefficient estimates are computed using the standardized design matrix $X_\bullet$ defined by

$$X_\bullet = X \begin{bmatrix} 1 & -\bar{X}_2 S_2^{-1} & -\bar{X}_3 S_3^{-1} & \dots & -\bar{X}_{p-1} S_{p-1}^{-1} & -\bar{X}_p S_p^{-1} \\ 0 & S_2^{-1} & 0 & \dots & 0 & 0 \\ & 0 & S_3^{-1} & & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ & & \vdots & & S_{p-1}^{-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & S_p^{-1} \end{bmatrix} = XT,$$

where $\bar{X}_j = n^{-1} \sum_{i=1}^n X_{ij}$ and $S_j^2 = (n-1)^{-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$ for $j \in \{2, \dots, p\}$. Let $\hat{\beta}_\bullet$ be the shrinkage estimator of the standardized coefficients. Since all the $S_j$'s will be positive, we invert $T$ to estimate the original $\beta$ with $T^{-1}\hat{\beta}_\bullet$. Our proposed fitted-value shrinkage procedures are invariant to this standardizing transformation of $X$.

We display side-by-side boxplots of the same-X losses from the 50 replications when $p = 75$ in Figure 1. Further numerical summaries of these results are in Table 4 in Section B.2 of the Supplementary material. Without surprise, our fitted-value shrinkage with oracle tuning **2n-Or** performed the best among these candidates. Our proposed estimator **2n-Es** and its two variants **2n-Es95** and **2n-Es90** followed and generally outperformed OLS, Ridge, and Lasso when $\tau \geq 10^{-0.5}$. Of the fitted-value shrinkage estimators, **2n-Es** outperformed **2n-Rep**. On the other hand, the modified thresholds **2n-Es90** and **2n-Es95** performed better than **2n-Es** for smaller values of $\tau$ (that correspond to smaller values of $\delta^2$). Also, **2n-G** outperformed Ridge, Lasso, **2n-ES**, and **2n-Es95** for smaller values of $\tau$. Ridge and Lasso slightly outperformed **2n-Es** when $\delta^2$ was small, but performed worse when $\delta^2$ was larger.

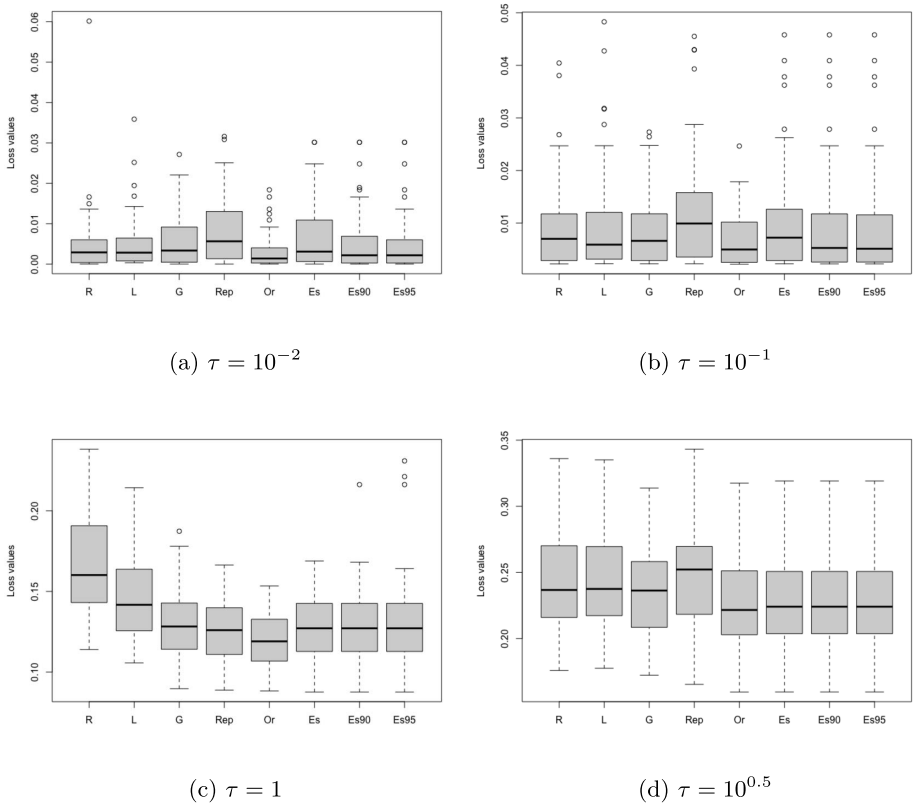
In Figure 8 in Section B.1 of the Supplementary material, we graphed the average same-X loss values over the 50 replications as a function of $\lambda$:

$$f_{\text{FVS}}(\lambda) = \|X\beta - X\hat{\beta}_\lambda\|^2/n, \tag{21}$$

$$f_{\text{Ridge}}(\lambda) = \|X\beta - X\hat{\beta}_{\text{Ridge};\lambda}\|^2/n, \tag{22}$$

where $\hat{\beta}_\lambda$ and $\hat{\beta}_{\text{Ridge};\lambda}$ are solutions for (2) and (19); and $\lambda \in \{10^{-7+0.01k} : j = 0, 1, \dots, 1150\}$. This allows a further comparison of fitted-value shrinkage (3) to Ridge regression (19). Further numerical summaries of these results are in Table 10 in Section B.2 of the Supplementary material. The minimum average same-X loss for fitted value shrinkage (2) was either less than or nearly equal to that of Ridge (19). However, the range of values of $\lambda$ that corresponded to average losses near the minima was much narrower for fitted-value shrinkage than it was for Ridge when medium to large values of $\tau$ were used.

We also display side-by-side boxplots of the observed same-X losses from the 50 replications when $p = 150$ in Figure 2 and when $p = 250$ in Figure 7 in



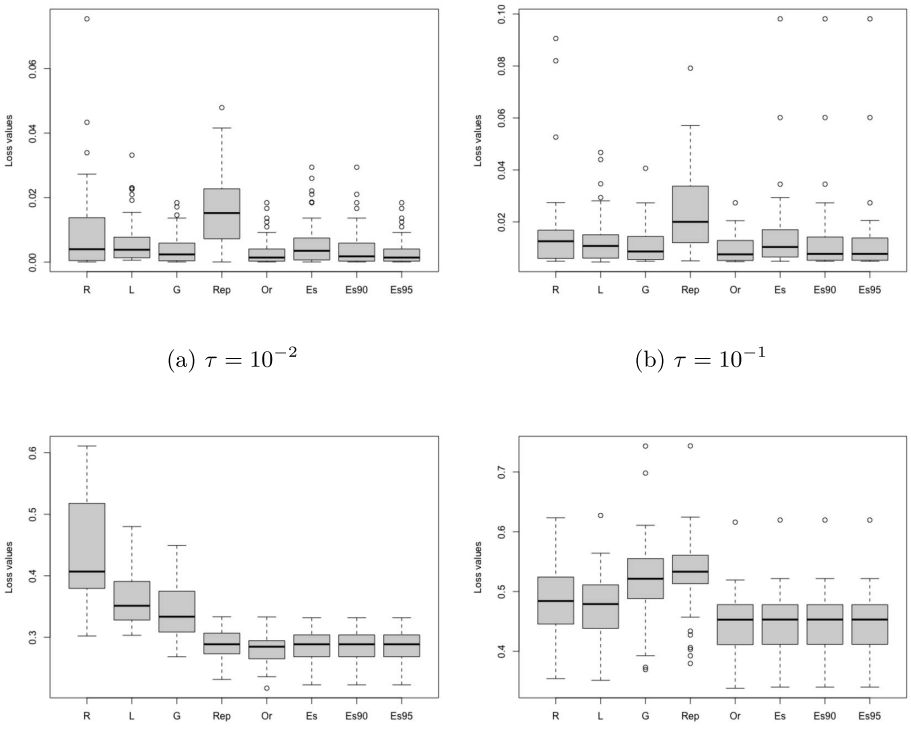

(a) $\tau = 10^{-2}$ (b) $\tau = 10^{-1}$

(c) $\tau = 1$ (d) $\tau = 10^{0.5}$

FIG 1. *Boxplots of the observed same-X loss values from the 50 replications when* $n = 300$ *and* $p = 75$. *We suppress "2n" from the third to the last estimators to simplify notation.*

Section B.1 of the Supplementary material. In addition, graphs of the average loss values as a function of $\lambda$ over $\lambda \in \{10^{-7+0.01k} : j = 0, 1, \ldots, 1100\}$ when $p = 150$ and $p = 250$ are in Figure 7 in Section B.1 of the Supplementary material. Further numerical summaries for this simulation with $p \in \{150, 250\}$ are in Table 5–6, and Table 11–12 in Section B.2 of the Supplementary material. These results with $p \in \{150, 250\}$ are similar to results when $p = 75$: the oracle method **2n-Or** was the best and our proposed estimators **2n-Es90**, **2n-Es95**, **2n-Es** were the most competitive when $\tau \geq 1$. However, the performance gap between our procedure with non-oracle tuning **2n-G** and our procedure using oracle tuning **2n-Es** when $\tau \geq 1$ has increased (Figure 2c, 2d). We expect this is related to the narrower valley observed in the graph of the average same-X loss as a function of the tuning parameter. As it was when $p = 75$, **2n-G** was the most competitive for smaller values of $\tau$.

In Tables 7–8 in Section B.2 of the Supplementary material, we report 95% simulation-based confidence intervals for the expected squared same-X loss difference (fitted value shrinkage minus Ridge). The confidence intervals are based on 50 replications for each pair of $(\tau, p)$. These differences were statistically in-




(a) $\tau = 10^{-2}$ (b) $\tau = 10^{-1}$

(c) $\tau = 1$ (d) $\tau = 10^{0.5}$

FIG 2. *Boxplots of the observed same-X loss values from the 50 replications when* $n = 300$ *and* $p = 150$. *See the caption of Figure 1 for more details.*

significant when $\tau$ was small, but were significant when $\tau$ was larger (with fitted value shrinkage outperforming Ridge). In Table 9 in Section B.2 of the Supplementary material, we also report the average realization of $|\hat{\gamma} - \gamma_{\text{opt}}|^2$ based on 100 independent replications from the same simulation setting. This quantity increased with $p$, but remained stable.

### 5.2. *Impact of invariance I: In the presence of categorical variables with different reference level coding schemes*

We generated $X \in \mathbb{R}^{n \times p}$ to have 25 numerical predictors and 3 categorical predictors, where each categorical predictor had 5 levels, and the 25th numerical predictor had interactions with the three categorical predictors. The design matrix $X$ had $p = 1 + 25 + 3 \times 4 \times 2 = 50$ columns including the intercept. The $n = 100$ observations of the 25 numerical predictors were independent draws from $N_{25}(0, \Sigma)$, where $\Sigma_{jk} = 0.5^{|j-k|}$. Observations of each categorical predictor were independently drawn from a 5-category multinomial distribution with equal category probabilities. As they were before, $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, 1)$.

We organize the regression coefficient vector as $\beta = (\beta_c, \beta_f) \in \mathbb{R}^{50}$, where $\beta_c$ corresponds to the first 26 columns of $X$, which have the numerical predictors; and $\beta_f$ corresponds to the remaining 24 columns of $X$, which have the categorical predictors and their interaction with the 25th numerical predictor. Let $X_c \in \mathbb{R}^{100\times 26}$ be the matrix with the first 26 columns of $X$. We generated $\beta_c = X_c^-(1_{26} + \tau_c Z_c)$; where $Z_c \sim N_{26}(0, I)$. We also generated

$$\begin{aligned}\beta_f &= (\tau_f X_f^- Z_f) * v,\\ v &= (w', w')' \in \mathbb{R}^{24},\\ w &= (1,2,0,0,0,2,1,0,0,0,2,1)' \in \mathbb{R}^{12},\end{aligned}$$

where $*$ denotes the elementwise product of two vectors of same dimension; $Z_f \sim N_{24}(0, I)$; $X_f \sim N_{24}(0, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$; $Z_c$ is independent of $Z_f$; and $X_c$ is independent of $X_f$. We consider $(\tau_c, \tau_f)$ in

$$\{(10^{-0.5}, 10^{-0.5}), (1, 10^{-0.5}), (10^{-0.5}, 1), (1,1), (10^{0.5}, 1), (1, 10^{0.5}), (10^{0.5}, 10^{0.5})\}.$$

The first 12 entries of $\beta_f$ correspond to the $3 * 4$ main effects columns for the 3 categorical predictors. The remaining 12 entries of $\beta_f$ correspond to interactions between the 3 categorical predictors and the 25th numerical predictor.

The first level of each categorical predictor was coded as the reference level in this design matrix $X$. We call this "Coding-1". To illustrate the effects of a coding change, we will also use a "Coding-2" design matrix $X_\bullet$, which is an invertible linear transformation of $X$ that uses the second, third, and fifth levels of the three categorical predictors as the reference levels for the first, second, and third categorical predictor, respectively.

The competitors were **2n-Es**, **2n-Es95**, **2n-Or**, **2n-G**, **2n-Rep**, **O**, **R**, **L**, and Group Lasso and its two variants. The Group lasso competitors were the following:

- **GL**: Group Lasso estimator proposed by Yuan and Lin [29].

$$\hat{\beta}_{\mathrm{GL}} = \underset{b\in\mathbb{R}^p}{\arg\min} \|Y - Xb\|^2 + \lambda \sum_{i=1}^{L} \sqrt{p_i}\|b^{(i)}\|_2,$$

  where $i \in \{1, \ldots, L\}$ is the index of each group except for one intercept column; $p_i$ is the size of the $i$-th group; and $b^{(i)}$ is the corresponding subvector.
- **MGL**: Modified Group Lasso proposed by Choi et al. [7].

$$\hat{\beta}_{\mathrm{MGL}} = \underset{b\in\mathbb{R}^p}{\arg\min} \|Y - Xb\|^2 + \lambda \sum_{i=1}^{L} \|b^{(i)}\|_2.$$

- **SGL**: Standardized Group Lasso proposed by Simon and Tibshirani [24].

$$\hat{\beta}_{\mathrm{SGL}} = \underset{b\in\mathbb{R}^p}{\arg\min} \|Y - Xb\|^2 + \lambda \sum_{i=1}^{L} \sqrt{p_i}\|X^{(i)}b^{(i)}\|_2,$$

  where $X^{(i)}$ is the $i$-th submatrix.

For the estimators that require tuning parameter selection, 10-fold cross validation was used. For **GL**, **MGL**, and **SGL**, we formed 31 groups: 25 single-member groups corresponding to the numerical predictors, 3 groups corresponding to the categorical predictors' main effects, and 3 groups corresponding to the interaction between the 25th numerical predictor and the three categorical predictors.

We display side-by-side boxplots of the same-X loss values from 50 replications in Figure 3. Further numerical summaries of these results are in Table 13 in Section B.3 of the Supplementary material. Except for $(\tau_c, \tau_f) = (10^{0.5}, 10^{0.5})$ and $(10^{0.5}, 10^{-0.5})$, **2n-Or** was the best in Coding-2. Generally, **2n-Es** was second-best in Coding-2. The methods **R**, **L**, **GL**, **MGL** and **SGL** generally performed well in Coding-1, which is not surprising considering the sparsity in the coefficient vector; however, in Coding-2, these methods were generally worse than **2n-Es**, which is invariant to this change in coding. This illustrates that the invariance of our proposed estimators is advantageous.

### *5.3. Impact of invariance II: General full-rank transformation of the model matrix (Low dimension)*

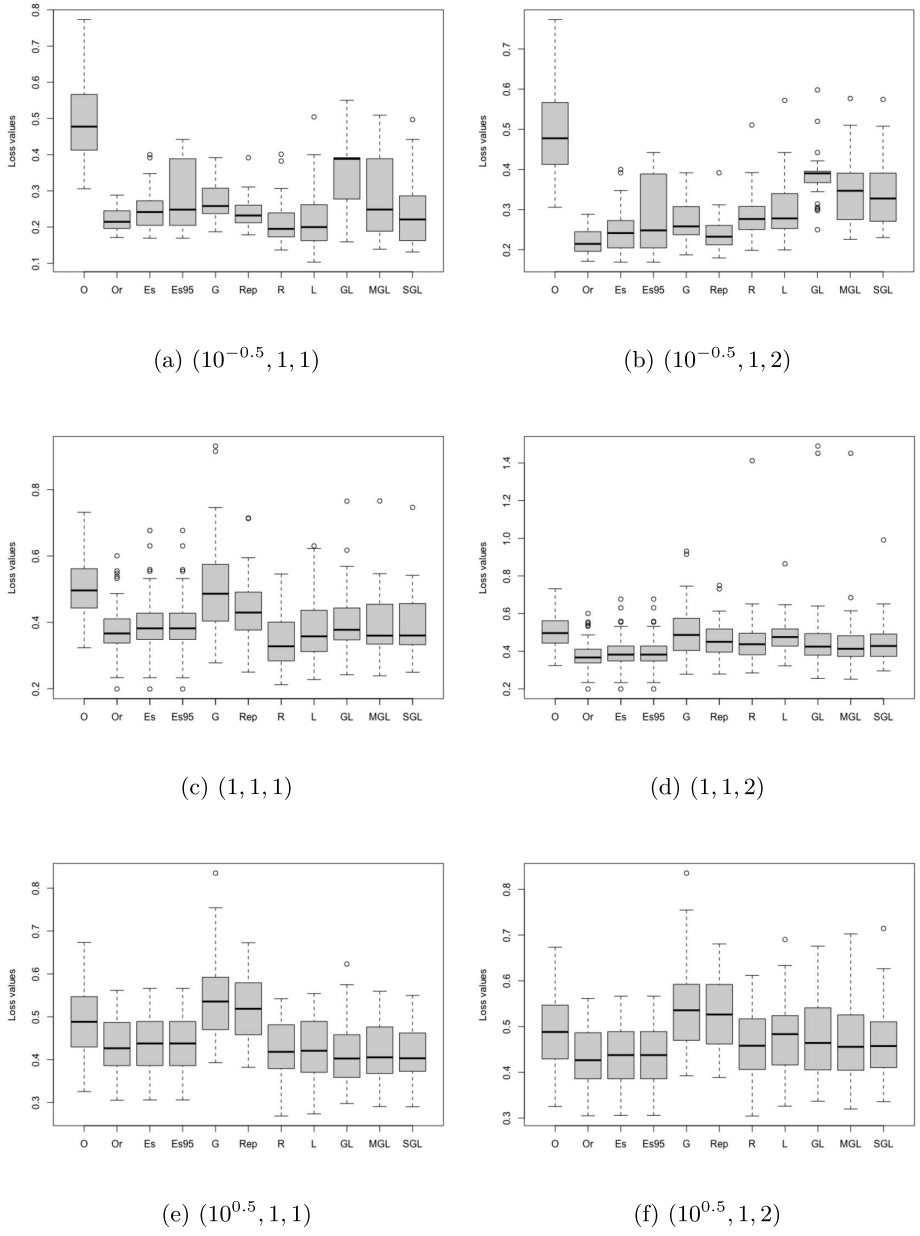
In this example, the design matrix $X \in \mathbb{R}^{n \times p}$ has ones in its first column and its $n$ row vectors (excluding the first entry) are drawn independently from $N_{p-1}(0, \Sigma)$ where $\Sigma_{jk} = 0.5^{|j-k|}$ and $(n, p) = (300, 150)$. We generated the regression coefficient vector $\beta$ as $\beta = u * v$, where each element of $u \in \mathbb{R}^p$ is an independent draw from the Uniform distribution on $(2^{-\psi-1}, 2^{-\psi})$ with $\psi \in \{0, 1\}$; and $v = (1, v_{-1}) \in \mathbb{R}^p$, where each element of $v_{-1} \in \mathbb{R}^{p-1}$ is an independent draw from Ber$(s)$ with $s \in \{0.025, 0.05, 0.1, 0.2, 0.3\}$. We generated $Y$ by (1).

We denote the model fitting scheme with this $X$ as "Coding-1". We also use a transformed design matrix $X_\bullet = XT$, where $T$ is a Gram-Schmidt orthogonalization of a matrix with all of its entries drawn independently from $N(0, 1)$. We refer the transformed design matrix $X_\bullet$ as "Coding-2".

In Figure 4, we display side-by-side boxplots of the observed same-X losses from the 50 replications from representative settings when $\psi = 0$. Further numerical summaries of the results are in Table 14 in Section B.4 of the Supplementary material. When $\tau = 1$, Lasso was the best in Coding-1 with Ridge following next. However, in Coding-2, **2n-Or** was the best and **2n-Es** was the second-best regardless of $\psi$ and $s$. Ridge and Lasso were significantly worse than our proposed estimator as well as OLS under Coding 2. This further illustrates the benefits of invariance.

### *5.4. High-dimensional experiments*

We used the same data generating model as in Section 5.1 except that $n = 200$, $p = 300$, and $\sigma \in \{2, 3\}$. Since **2n-Es** is not applicable in high dimensions, we tested variants of **2n-G** that used 5, 10, and $n$-fold cross validation. However, except for a few cases, the number of folds used for tuning parameter selection



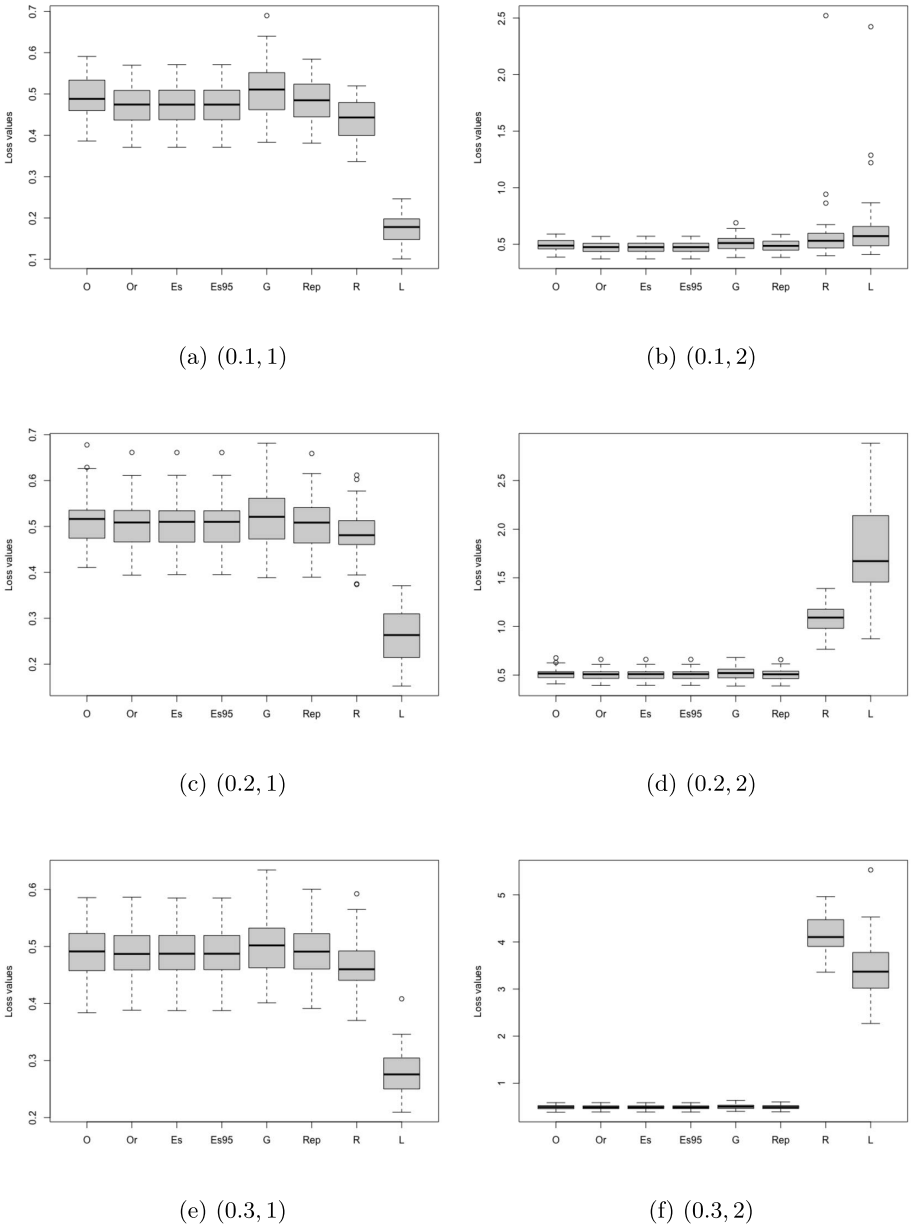

(a) $(10^{-0.5}, 1, 1)$ (b) $(10^{-0.5}, 1, 2)$

(c) $(1, 1, 1)$ (d) $(1, 1, 2)$

(e) $(10^{0.5}, 1, 1)$ (f) $(10^{0.5}, 1, 2)$

FIG 3. *Boxplots of the observed same-X loss values from the 50 replications in the simulation settings of Section 5.2. We suppress "2n" from the third to the last estimators for notational simplicity. Each plot is labeled with the* $(\tau_c, \tau_f, \text{Coding number})$ *value used.*

for **2n-G** did not have a significant impact on the performance. So we only present results from 10-fold cross validation with the same label **2n-G** as in




(a) $(0.1, 1)$ (b) $(0.1, 2)$

(c) $(0.2, 1)$ (d) $(0.2, 2)$

(e) $(0.3, 1)$ (f) $(0.3, 2)$

FIG 4. *Boxplots of the observed same-X loss values from the 50 replications in the simulation settings of Section 5.3 when $\psi = 0$. We suppress "2n" from the third to the last estimators for notational simplicity. Each plot is labeled with the* $(s, \text{Coding number})$ *value used.*

the previous sections. In addition, we tried different $\alpha$ values that control the matrix $K$ in (16) for the following variants of **2n-Rep**:

- **2n-Rep1**: Same as original **2n-Rep** with $\alpha = n(2\|Y - P_1 Y\|^2)^{-1}$.
- **2n-Rep2**: $\bar{\gamma}_c$ (16) using $\check{\sigma}^2$ with $\alpha = n^{3/2}(2\|Y - P_1 Y\|^2)^{-1}$.
- **2n-Rep3**: Same as original **2n-Rep** with $\alpha = n^2(2\|Y - P_1 Y\|^2)^{-1}$.
- **2n-Rep4**: Same as original **2n-Rep** with $\alpha = n^3(2\|Y - P_1 Y\|^2)^{-1}$.

**2n-Rep1**, **2n-Rep3**, and **2n-Rep4** are based on our variance estimator (15); and **2n-Rep2** is based on the variance estimator proposed by Liu et al. [22]. From Proposition 4, **2n-Rep1** is suitable when $\delta^2 = \|\mu - P_1\mu\|^2$ is large relative to $n\sigma^2$, e.g. $\gamma_{\text{opt}} \to 1$. In contrast, based on Proposition 4 and 5, **2n-Rep2**, **2n-Rep3**, and **2n-Rep4** are suitable when $\delta^2 = \|\mu - P_1\mu\|^2$ is small relative to $n\sigma^2$, e.g. $\gamma_{\text{opt}} \to 0$. As we will illustrate below, **2n-Rep3** generally outperforms **2n-Rep2** and **2n-Rep4**. So we recommend using **2n-Rep1** for stronger signals and **2n-Rep3** for weaker signals.

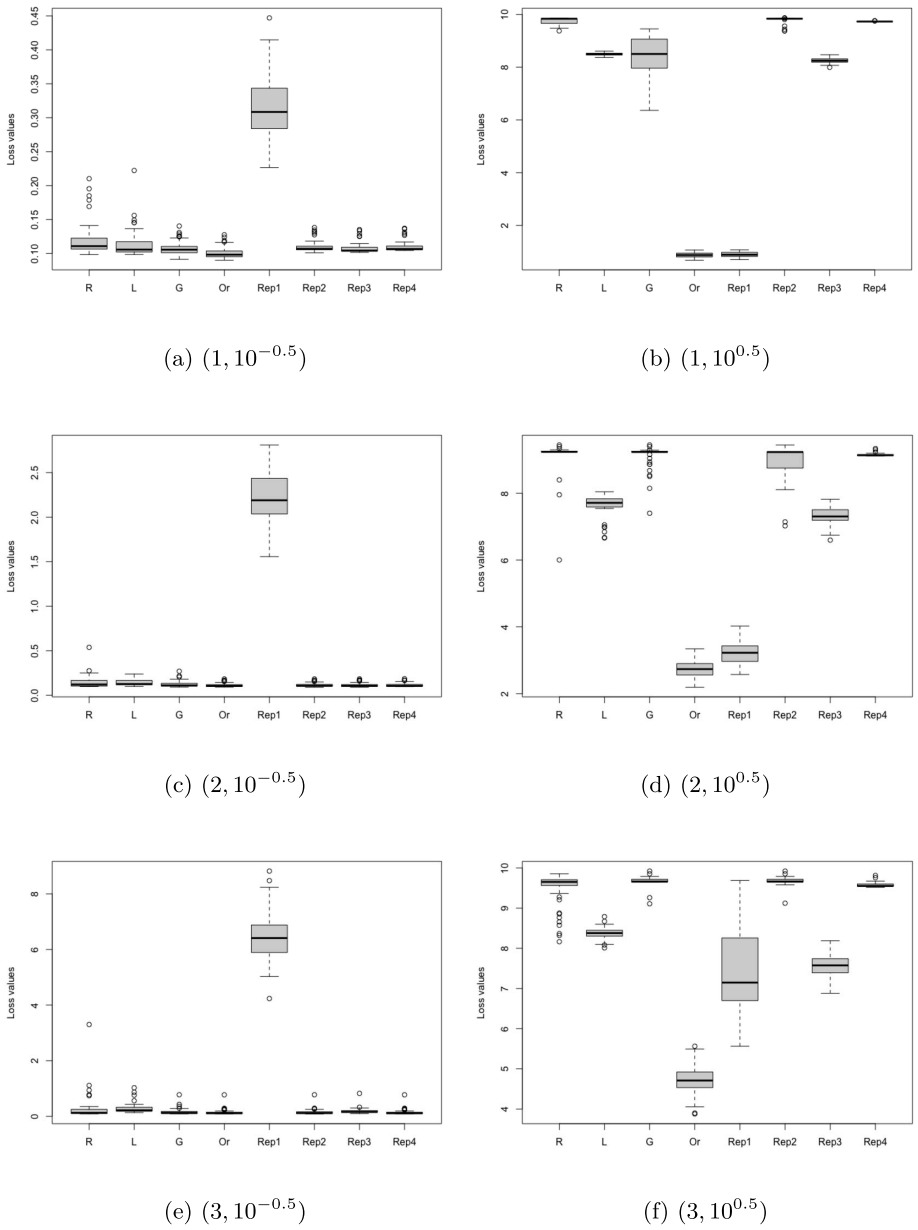
In Figure 5, we display side-by-side boxplots of the observed same-X losses from the 50 replications when $n = 200$ and $p = 300$. There are additional boxplots displayed in Figure 11 in Section B.1 of the Supplementary material. Further numerical summaries of the results are in Table 15–17 in Section B.5 of the Supplementary material. In general, when $\delta^2 = \|\mu - P_1\mu\|^2$ was large relative to $n\sigma^2$, which corresponds to the situation $\gamma_{\text{opt}} \to 1$, **2n-Rep1** performed substantially better than Ridge, Lasso, **2n-G**, **2n-Rep2**, and **2n-Rep4** (Figure 5b, 5d). The method **2n-Rep1** also outperformed **2n-Rep3**, but the same-X prediction difference was relatively smaller than the others. Furthermore, larger $\delta^2$ led to improved tuning-parameter selection for **2n-Rep1**. However, **2n-G**, Ridge, and Lasso performed better when $\delta^2$ was small relative to $n\sigma^2$. In this setting, which corresponds to $\gamma_{\text{opt}} \to 0$, **2n-Rep2**, **2n-Rep3**, and **2n-Rep4** performed as well as or better than Ridge, Lasso, and **2n-G** (see Figure 5a, 5c, 5e). On the other hand, **2n-Rep1** struggled for this case. The method **2n-Rep3** was relatively more stable than the other **2n-Rep** versions, and it consistently outperformed Ridge and Lasso regardless of $\delta^2$.

In Figure 12 (see Section B.1 of the Supplementary material), we display average loss values over the 50 replications as a function of $\lambda$ over $\lambda \in \{10^{-7+0.01k} : j = 0, 1, \ldots, 1100\}$ when $n = 200$ and $p = 300$ (see (21) and (22)). Further numerical summaries of these results are in Table 18 in Section b.5 of the Supplementary material. These results look similar to lower-dimensional results displayed in Figures 8–10 (see Section B.1 of the Supplementary material), except the curve valleys are narrower for fitted-value shrinkage.

### 5.5. *Impact of (approximate) invariance III: General full-rank transformation of the design matrix (High dimension)*

We again generated $X \in \mathbb{R}^{n \times p}$ to have ones in its first column and and its $n$ row vectors (excluding the first entry) were drawn independently from $N_{p-1}(0, \Sigma)$ where $\Sigma_{jk} = 0.5^{|j-k|}$ and $(n, p) = (200, 300)$. We used (1) again for the data generating procedure, where we consider $\sigma \in \{1, 3\}$.

We generated $\beta = u * v$, where each element of $u \in \mathbb{R}^p$ is an independent draw from the Uniform distribution on $(1/4, 1/2)$ when $\sigma = 1$; and on $(1/6, 1/3)$ when




(a) $(1, 10^{-0.5})$ (b) $(1, 10^{0.5})$

(c) $(2, 10^{-0.5})$ (d) $(2, 10^{0.5})$

(e) $(3, 10^{-0.5})$ (f) $(3, 10^{0.5})$

FIG 5. *Boxplots of the observed same-X loss values from the 50 replications when* $(n, p) = (200, 300)$. *Each plot is labeled with the* $(\sigma, \tau)$ *value used.*

$\sigma = 3$; $v = (1, v_{-1}) \in \mathbb{R}^p$, where each element of $v_{-1} \in \mathbb{R}^{p-1}$ is an independent draw from $\text{Ber}(s)$. When $\sigma = 1$, we vary $s \in \{0.025, 0.05, 0.1, 0.2, 0.3\}$, and consider **2n-Rep1** for the proposed estimator. On the other hand, when $\sigma = 3$,

we vary $s \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$, and consider **2n-Rep3** for the proposed estimator as well as **2n-Rep2**, which is known to be efficient in the same setting that **2n-Rep3** is (see Proposition 4). We dropped **2n-Rep4**, since it showed relatively unstable performance in comparison to **2n-Rep3** in Section 5.4.

As in Section 5.3, we denote the model fitting scheme with this $X$ as "Coding-1". We also use a transformed design matrix $X_\bullet = XT$, where $T$ through a Gram-Schmidt orthogonalization of a matrix with all of its entries drawn independently from $N(0, 1)$. We refer the transformed design matrix $X_\bullet$ as "Coding-2".

In Figure 6, we display side-by-side boxplots of the observed same-X losses from the 50 replications in few representative settings. Further numerical summaries of the simulation results are in Tables 19–20 in Section B.6 of the Supplementary matrial. When $\tau = 1$, Lasso was generally the best in Coding-1 and Ridge was second best. However, in Coding-2, **2n-Or** was the best and **2n-Rep1** was second best, regardless of $\psi$ and $s$. Ridge and Lasso performed significantly worse than **2n-Rep1** after transformation.

On the other hand, when $\sigma = 3$, **2n-Or** achieved the best performance for both Coding 1 and 2 except for $s = 0.05$ where Ridge was the best for both Coding 1 and 2. For $s \in \{0.01, 0.02, 0.03, 0.04\}$, **2n-Rep3** was competitive, so were **2n-G** and **2n-Rep2**, which is similar to the pattern observed in Section 5.4 when $\delta^2$ was low.

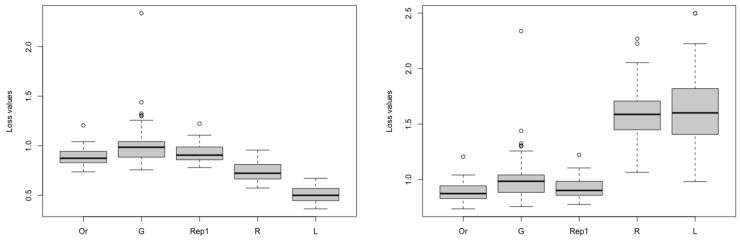
Although the **2n-Rep** methods lack exact invariance in high dimensions, their performance difference between Coding 1 and Coding 2 was not significant.

## 6. Data examples

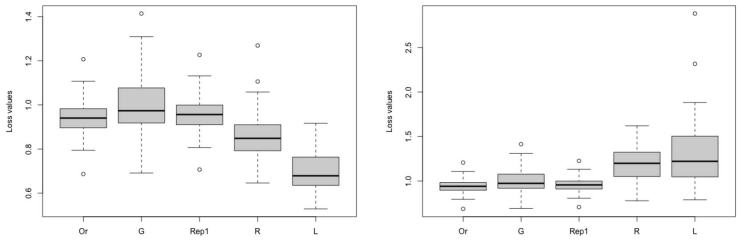
### *6.1. Low dimensional data experiments*

We compared our proposed fitted-value shrinkage procedures to competitors on three data sets. We used the same non-oracle estimators as the previous section except we excluded **2n-Es90** because it performed similarly to **2n-Es95**. Each data example was analyzed using the following procedure: For 50 independent replications, we randomly selected 70% of the subjects for the training set and used the remaining subjects as the test set. Tuning parameter selection was done using the training set and prediction performance was measured using squared error loss on the test set. The following is the short description of the three low-dimensional data set we examined.

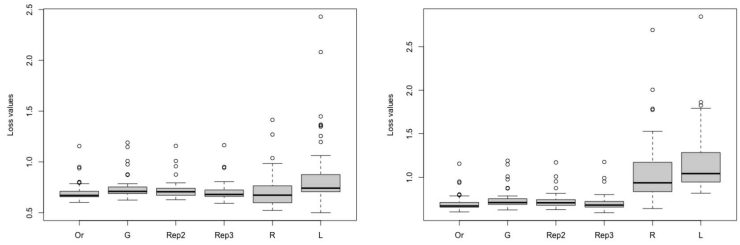
(FF): The Forest Fire (FF) data are from from Cortez and Morais [10] and are stored at the UCI Machine learning repository via `https://archive.ics.uci.edu/dataset/162/forest+fires`. There are 517 observations corresponding to forest fires in Portugal from 2000 to 2003. The response is the total burned **area** (in $ha$) from the fire, which was transformed with $x \mapsto \ln(x + 1)$, which was suggested by Cortez and Morais [10]. There were originally 13 attributes. However, in the pre-processing step, since we are not focusing on spatio-temporal methods, we excluded time, date and location coordinates. After this processing, the full-data



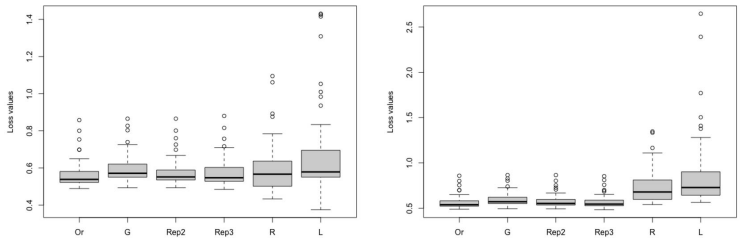

(a) $(1, 0.1, 1)$ (b) $(1, 0.1, 2)$




(c) $(1, 0.2, 1)$ (d) $(1, 0.2, 2)$




(e) $(3, 0.03, 1)$ (f) $(3, 0.03, 2)$




(g) $(3, 0.04, 1)$ (h) $(3, 0.04, 2)$

FIG 6. *Boxplots of the observed same-X loss values from the 50 replications in the simulation settings of Section 5.5. We suppress "2n" from the third to the last estimators for notational simplicity. Each plot is labeled with the* $(\sigma, s,$ Coding number$)$ *value used.*

design matrix had $(n, p) = (517, 9)$ with 8 numerical-variable columns and one intercept column.

(GDP): The GDP data (GDP) are from Barro and Lee [4]. These data consist of 161 observations of GDP growth rates for the two periods 1965–1975 and 1975–1985. The data are also in the R package `quantreg` [20]. The response is **Annual change per capita GDP**. There are 13 numerical predictors, e.g. Initial per capita GDP, Life expectancy. We also added a quadratic term for the predictor **Black Market Premium**. After processing, the full-data design matrix has $(n, p) = (161, 15)$.

(FC): The Forecast data set (FC) is from Cho et al. [6] for the purpose of bias correction for the Local Data Assimilation and Prediction System (LDAPS), which is a numerical weather report model used by Korea Administration (KMA), Seoul, South Korea. It has a public access through https://archive.ics.uci.edu/dataset/514/bias+correction+of+numerical+prediction+model+temperature+forecast.
The data are regional observations from 2013 to 2017, from which we randomly selected 500. We used the true maximal temperature of the next day as the response and removed date, station ID, and true minimal temperature of the next day. The full-data design matrix had $(n, p) = (500, 20)$.

In Table 1, we display mean squared prediction errors averaged over 50 training/test set splits for the three data examples (FF), (GDP), and (FC). Our fitted-value shrinkage estimators performed similarly to Ridge and Lasso, which both lack invariance to invertible linear transformations of the design matrix.

TABLE 1

*The performance comparison table for three data examples in Section 6.1. The values are the mean squared prediction errors averaged over 50 training/test set splits. The numbers in parentheses are normalized sample standard deviations. The column labels are defined in Section 5.1. Boldface indicates the best model. Underlined is our main proposed estimator.*

| Performance comparison table | | | | | | | |
|---|---|---|---|---|---|---|---|
| Data set | 2n-Rep | 2n-G | 2n-Es | 2n-Es95 | OLS | Ridge | LASSO |
| Forest Fire (FF) | 2.0031 (0.0301) | 2.0160 (0.0316) | <u>2.0017</u> (0.0300) | **1.9792** (0.0293) | 2.1834 (0.0551) | 2.0189 (0.0322) | 2.0572 (0.0411) |
| GDP growth (GDP) | **3.106e-04** (7.465e-06) | 3.125e-04 (7.615e-06) | <u>3.139e-04</u> (7.706e-06) | 3.139e-04 (7.706e-06) | 3.231e-04 (8.120e-06) | 3.112e-04 (8.126e-06) | 3.149e-04 (8.307e-06) |
| Forecast (FC) | 1.9045 (0.0327) | 1.8997 (0.0325) | <u>1.9028</u> (0.0327) | **1.8933** (0.0318) | 2.0489 (0.0371) | 1.8990 (0.0321) | 1.8978 (0.0318) |

### 6.2. *Low dimensional data analyses with categorical variables and their interactions*

We analyzed two data sets from existing R packages to illustrate the performance of our estimators when categorical variables with interactions are present in the model. The competitors and setup for the data experiments are nearly identical to the previous Section 5.2, except we added three new fitted-value shrinkage estimators that shrink toward the submodel without interactions instead of the intercept-only model. We refer the readers to Section C of the

Supplementary material for the definition of the submodel shrinkage. These new submodel shrinkage methods are labeled **2n-Repsb**, **2n-Gsb**, **2n-Essb**, **2n-Es95sb**, and they respectively correspond to **2n-Rep**, **2n-G**, **2n-Es**, and **2n-Es95**.

The following is a description of the data examples:

(Dia-1): The Diamonds data set is from Diamonds data frame in the R package `Stat2Data` [5], and it was obtained from `https://awesomegems.com/`. The are $n = 351$ subjects and the response is the price of the diamond (in dollars). We divided price per 1000 and used it for the response variable. We further removed total price from the predictors. There are 2 categorical predictors: color (with levels D to J) and clarity (with levels IF, VVS1, VVS2, VS1, VS2, SI1, SI2, and SI3). We divided color into 5 levels (D, E, F, G, and (H, I, J)), and categorized the clarity into 3 levels ((IF, VVS1, VVS2), (VS1, VS2), (SI1, SI2, SI3)). We used reference-level coding in the design matrix, where (H, I, J) was the reference level for color; and (SI1, SI2, SI3) was the reference level for clarity. Interactions between color and depth as well as clarity and depth were added. The full design matrix has $(n, p) = (351, 15)$ and the submodel with linear terms only has $p = 9$.

(Dia-2): The setting is identical to that of (Dia-1), except we used the category (VS1, VS2) as the reference level for coding the categorical predictor clarity.

(NG-1): The NaturalGas data is from Baltagi [3], and is in the R package `AER` [19]. There are 138 observations on 10 variables. We removed state name and year and added an interaction between state code and heating degree days. We set the response as consumption divided by 10000. The reference level for state code, which is the only categorical predictor, was set to 35 (NY). The full-data design matrix had $(n, p) = (138, 17)$ and the submodel without interactions had $p = 12$.

(NG-2): This is the same as (NG-1), except the reference level for state code was set to 5 (CA).

We display mean squared prediction errors averaged over 50 training/test set splits for these data examples in Table 2. Our proposed estimators performed similarly or better than Ridge and Lasso. We also notice that changing the way that categorical predictors were encoded in the design matrix changes the performance of Ridge and Lasso, which lack invariance. Furthermore, Group Lasso and its two standardized versions had unstable performance when the base level coding was changed, which does not happen to our invariant methods.

### *6.3. High dimensional data experiments*

For high-dimensional data examples, we randomly selected subjects from existing data sets so that there were fewer subjects than predictors. We used the same splitting and evaluation procedure that we used in Sections 6.1 and 6.2.

TABLE 2

*The performance comparison table for two data sets in Section 6.2 with two different coding strategies. The values are mean squared prediction errors averaged over 50 training/test splits. The numbers in parentheses are normalized sample standard deviations. The labels are defined in Section 5.1 except that last three are illustrated in Section 6.2. The suffix "sb" for each estimator corresponds to shrinking towards the submodel without interactions instead of shrinking towards the intercept-only model (see Section C for details). Boldface indicates the best model. Underlined are our main estimator and its submodel shrinkage variant.*

| Performance comparison table | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data set | 2n-Rep | 2n-Repsb | 2n-G | 2n-Gsb | 2n-Es | 2n-Essb | 2n-Es95 | 2n-Es95sb | OLS | Ridge | LASSO | GL | MGL | SGL |
| Diamonds (Dia-1) | 1.2538 (0.0303) | 1.2557 (0.0300) | 1.2545 (0.0298) | **1.2501** (0.0293) | 1.2557 (0.0295) | 1.2507 (0.0294) | 1.2557 (0.0295) | 1.2599 (0.0294) | 1.2587 (0.0293) | 1.2668 (0.0297) | 1.2675 (0.0297) | 1.2699 (0.0290) | 1.2693 (0.0290) | 1.2569 (0.0301) |
| Diamonds (Dia-2) | 1.2538 (0.0303) | 1.2557 (0.0300) | 1.2545 (0.0298) | **1.2501** (0.0293) | 1.2557 (0.0295) | 1.2507 (0.0294) | 1.2557 (0.0295) | 1.2599 (0.0294) | 1.2587 (0.0293) | 1.2615 (0.0296) | 1.2624 (0.0299) | 1.2697 (0.0286) | 1.2677 (0.0287) | 1.2592 (0.0296) |
| Natural gas (NG-1) | 5.2558 (0.2285) | 5.2485 (0.2262) | 5.2899 (0.2282) | 5.2711 (0.2217) | 5.2559 (0.2285) | 5.2523 (0.2257) | 5.2559 (0.2285) | 5.2523 (0.2257) | 5.2545 (0.2279) | 5.2913 (0.2195) | **5.1254** (0.2285) | 5.2972 (0.2241) | 5.3010 (0.2242) | 5.1584 (0.2216) |
| Natural gas (NG-2) | 5.2558 (0.2285) | 5.2486 (0.2262) | 5.2899 (0.2282) | 5.2711 (0.2217) | 5.2559 (0.2285) | 5.2523 (0.2257) | 5.2559 (0.2285) | 5.2523 (0.2257) | 5.2545 (0.2279) | 5.3563 (0.2195) | 5.2889 (0.2285) | 5.2261 (0.2171) | **5.2010** (0.2168) | 6.7549 (0.2877) |

The competitors are same as those considered in Section 5.4 except that we excluded **2n-Rep4** which had nearly identical performance to **2n-Rep3** in the simulation study. The following is a description of the examples:

(mtp): The data set mtp comes from Karthikeyan et al. [18] and is available at the OpenML repository via https://www.openml.org/search?type=data&status=\active&id=405. There are 4450 subjects with 203 numerical measurements. The response is oz203. We randomly selected 120 subjects and removed the 23 predictors that had fewer than 30 distinct values, which ensured that there were no constant columns in the 120-row design matrix other than the intercept column. The full-data design matrix had $(n, p) = (120, 180)$.

(topo): The topo.2.1 data set is from Feng et al. [12] and is available through the OpenML repository via https://www.openml.org/search?type=data&sort=runs&id=422\&status=active. There are 8885 subjects with 267 numerical measurements. The response is oz267. We randomly selected 180 subjects and removed the 22 predictors that had fewer than 30 distinct values. After this, there were 34 constant columns (other than the intercept) that were also removed. The R code for this processing is in Section D.0.1 of the Supplementary material. The full-data design matrix has $(n, p) = (180, 214)$.

(tecator): The tecator data set comes from Thodberg [25], and is also available from OpenML repository via https://www.openml.org/search?type=data&status=active&sort=\runs&order=desc&id=505. We randomly selected 100 subjects and removed 22 principal components. The response is fat content. The full-data design matrix had $(n, p) = (100, 103)$.

In Table 3, we report the mean squared prediction errors averaged over 50 training/test set splits for these data examples. We see that **2n-Rep2**, **2n-Rep3** had similar prediction performance compared to **2n-Rep1**. In contrast to its same-X loss performance in simulations, the cross validation version of our method **2n-G** gave reasonable out-of-sample prediction performance. Generally,

**2n-Rep1** and **2n-Rep2** performed competitively compared with Ridge and Lasso, and **2n-Rep3** followed the next within **2n-Rep**s.

TABLE 3

*The performance comparison table for three high dimensional real data sets in Section 6.3. The values are mean squared prediction errors averaged over 50 replications. The numbers in parentheses are normalized sample standard deviations. The labels are defined in Section 5.4. Boldface indicates the best model.*

| Performance comparison table | | | | | | | |
|---|---|---|---|---|---|---|---|
| Data set | OLS | 2n-G | Ridge | LASSO | 2n-Rep1 | 2n-Rep2 | 2n-Rep3 |
| mtp | 0.5075 | 0.0266 | 0.0262 | **0.0227** | 0.0257 | 0.0261 | 0.0258 |
| (mtp) | (0.0592) | (0.0010) | (0.0027) | (0.0017) | (0.0008) | (0.0009) | (0.0008) |
| topo.2.1 | 0.07524 | 0.00085 | 0.00096 | 0.00087 | 0.00084 | **0.00083** | 0.00084 |
| (topo) | (2.72e-02) | (2.75e-05) | (7.71e-05) | (3.27e-05) | (3.08e-05) | (2.73e-05) | (3.08e-05) |
| tecator | 2.4400 | 2.4592 | 2.7251 | 2.7120 | **2.4266** | 2.4318 | 2.4733 |
| (tecator) | (0.2359) | (0.2348) | (0.1913) | (0.1940) | (0.2202) | (0.2188) | (0.2148) |

## 7. Discussion

Lasso and ridge-penalized least squares are popular and powerful in practice. However, their fitted values lack invariance to invertible linear transformations of the design matrix, which is undesirable when there are categorical predictors and interactions. Our simulation studies and data analyses illustrated that our proposed method performed comparably to ridge-penalized least squares, and so we recommend that practitioners use our method in any problem that they would use ridge-penalized least squares. Our method serves as a companion to ridge-penalized least squares, with the advantage of preserving invariance to invertible linear transformations of the design matrix.

The fitted-value shrinkage idea presented here to fit linear regression models with invariance can be extended to more complicated settings. For example, one could fit a logistic regression model by minimizing the negative loglikelihood plus the penalty $\lambda\|X\beta - \hat{w}1_n\|^2$, where $\hat{w}$ is the sample log-odds that the response takes its first category. We are currently developing this procedure.

A Bayesian formulation of our method may also be interesting. In addition, one could study methods that combine our proposed invariant shrinkage penalty with regular shrinkage penalties like the lasso or ridge.

## Acknowledgement

The authors thank the editor, associate editor, and referees for their helpful comments and suggestions.

## Supplementary Material

**Supplement to "Fitted value shrinkage"**
(doi: 10.1214/24-EJS2303SUPP; .pdf).

## References


[1] Azriel, D. (2019). The conditionality principle in high-dimensional regression. *Biometrika* **106**(3), 702–707. MR3992399

[2] Azriel, D. and Schwartzman, A. (2020). Estimation of linear projections of non-sparse coefficients in high-dimensional regression. *Electronic Journal of Statistics* **14**(1), 174 – 206. MR4047998

[3] Baltagi, B. H. (2002). *Econometrics*. Heidelberg: Springer Berlin. MR2814522

[4] Barro, R. and Lee, J. (1994). Data set for a panel of 138 countries. *discussion paper, NBER* **138**.

[5] Cannon, A., Cobb, G., Hartlaub, B., Legler, J., Lock, R., Moore, T., Rossman, A., and Witmer, J. (2019). Stat2data: Datasets for stat2. https://CRAN.R-project.org/package=Stat2Data. R package version 2.0.0.

[6] Cho, D., Yoo, C., Im, J., and Cha, D. (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and space science* **7**(4).

[7] Choi, Y., Park, R., and Seo, M. (2012). Lasso on categorical data.

[8] Cook, R. D., Forzani, L., and Rothman, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics* **7**, 3059–3088. MR3151762

[9] Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research* **6**(2), 167–183. PMID: 9261914.

[10] Cortez, P. and Morais, A. (2007). Efficient forest fire occurrence prediction for developing countries using two weather parameters. *Environmental Science, Computer Science*.

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**(456), 1348–1360. MR1946581

[12] Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S., and Young, S. S. (2003). Predictive toxicology: benchmarking molecular descriptors and statistical methods. *Journal of Chemical Information and Computer Sciences* **43**(5), 1463–1470. PMID: 14502479.

[13] Ham, D. and Rothman A. J. (2024) Supplement to "Fitted value shrinkage" https://doi.org/10.1214/24-EJS2303

[14] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**(2), 109–135.

[15] Hastie, T. and Tibshirani, R. (2004, 07). Efficient quadratic regularization for expression arrays. *Biostatistics* **5**(3), 329–340.

[16] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67. MR4165988

[17] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417–441.

[18] Karthikeyan, M., Glen, R. C., and Bender, A. (2005). General melting

point prediction based on a diverse compound data set and artificial neural networks. *Journal of Chemical Information and Modeling* **45**(3), 581–590. PMID: 15921448.

[19] Kleiber, C. and Zeileis, A. (2008). Applied econometrics with R. https://CRAN.R-project.org/package=AER. ISBN 978-0-387-77316-2.

[20] Koenker, R. (2022). quantreg: Quantile regression. https://CRAN.R-project.org/package=quantreg. R package version 5.94. MR2268657

[21] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338. MR1805785

[22] Liu, X., rong Zheng, S., and Feng, X. (2020). Estimation of error variance via ridge regression. *Biometrika* **107**, 481–488. MR4108940

[23] Rosset, S. and Tibshirani, R. (2020). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association* **115**(529), 138–151. MR4078450

[24] Simon, N. and Tibshirani, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica* **22**(3), 983. MR2987480

[25] Thodberg, H. H. (2015). Tecator meat sample dataset. http://lib.stat.cmu.edu/datasets/tecator.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B* **58**, 267–288. MR1379242

[27] Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah (Ed.), *Multivariate Analysis*, pp. 391–420. New York: Academic Press. MR0220397

[28] Xie, W. Z. (1988). A simple way of computing the inverse moments of a non-central chi-square random variable. *Journal of econometrics* **37**(3), 389–393. MR0936546

[29] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67. MR2212574

[30] Zhang, A. R. and Zhou, Y. (2020). On the non-asymptotic and sharp lower tail bounds of random variables. *Stat* **9**(1), e314. MR4193419

[31] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**(2), 894–942. MR2604701

[32] Zhao, J., Zhou, Y., and Liu, Y. (2023). Estimation of linear functionals in high-dimensional linear models: From sparsity to nonsparsity. *Journal of the American Statistical Association* **0**(0), 1–13. MR4766011

[33] Zhu, Y. (2020). A convex optimization formulation for multivariate regression. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 17652–17661. Curran Associates, Inc.

[34] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**(476), 1418–1429. MR2279469