Contemporary Mathematics
Volume **622**, 2014
http://dx.doi.org/10.1090/conm/622/12438

Geometry and properties of generalized ridge regression in high dimensions

Hemant Ishwaran and J. Sunil Rao

Abstract. Hoerl and Kennard proposed generalized ridge regression (GRR) over forty years ago as a means to overcome deficiencies in least squares in multicollinear problems. Because high-dimensional regression naturally involves correlated predictors, in part due to the nature of the data and in part due to artifact of the dimensionality, it is reasonable to consider GRR for such problems. We study GRR when the number of predictors p exceeds the sample size n. A novel geometric interpretation for GRR is described in terms of a uniquely defined least squares estimator and lends insight into its properties. It is shown that GRR possesses a shrinkage property useful in correlated settings and that in sparse high-dimensional settings it can have excellent performance but no such guarantees hold in non-sparse settings. We describe a computationally efficient representation for GRR requiring only a linear number of operations in p, thus making GRR computationally applicable to high dimensions.

1. Introduction

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ be a response vector and $\mathbf{X} = (\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(p)})$ an $n \times p$ design matrix, where $\mathbf{X}_{(k)} = (x_{k,1}, \ldots, x_{k,n})^T$ denotes the kth column of $\mathbf{X}$. It is assumed that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_0^2 > 0$. The true value for the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is unknown and is denoted by $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$. In this paper, we focus on linear regression models (1.1) in high-dimensional scenarios where $p > n$.

The so-called "big-p small-n problem" poses unique obstacles for estimation of $\boldsymbol{\beta}_0$. One significant concern is multicollinearity. With very large p, the sample correlation between variables can become sizeable as a pure artifact of the dimensionality. Groups of variables become highly correlated with other groups of variables sporadically. These effects can even occur when the population design matrix is orthogonal, i.e., $\mathbb{E}(\mathbf{X}_{(k)}^T \mathbf{X}_{(j)}) = 0$ if $k \neq j$ (see [**1**, **3**] for a discussion of these points). Multicollinearity is further compounded as variables collected in high-dimensional applications are often naturally correlated because of the underlying technology

2010 *Mathematics Subject Classification.* Primary 62G99.
The first author was supported by DMS grant 0705037 from the National Science Foundation.

or science: for example, gene expression values obtained from DNA microarrays, or genotype data collected using SNP arrays in Genome Wide Association Studies (GWAS).

Over 40 years ago, Hoerl and Kennard [**6**, **7**] proposed generalized ridge regression (GRR), a method specifically designed for correlated and ill-conditioned settings. Although it is unlikely they envisioned using GRR in problems where p could be orders of magnitudes larger than n, it is natural to wonder if it can be applied effectively in such contexts.

We recall the definition of GRR. Let $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_k\}_{k=1}^p$ be a $p \times p$ diagonal matrix with diagonal entries $\lambda_k > 0$. The GRR estimator with ridge matrix $\mathbf{\Lambda}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = (\mathbf{Q} + \mathbf{\Lambda})^{-1}\mathbf{X}^T\mathbf{Y},$$

where $\mathbf{Q} = \mathbf{X}^T\mathbf{X}$. An important property of GRR is that $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ is well defined even when $\mathbf{Q}$ is non-invertible. An alternative representation for $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ is in terms of ℓ_2-penalization:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{k=1}^p \lambda_k\beta_k^2\right\}, \tag{1.2}$$

where $\|\cdot\|$ is the ℓ_2-norm. Setting $\mathbf{\Lambda} = \lambda\mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix, yields the ridge estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{Q}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$ as a special case. The parameter $\lambda > 0$ is referred to as the ridge parameter. Setting $\lambda = 0$ and assuming that $\mathbf{Q}$ is invertible yields the OLS (ordinary least squares) estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{Q}^{-1}\mathbf{X}^T\mathbf{Y}$.

Much of the recent effort to address high-dimensional problems has focused on ℓ_1-penalization (lasso) methods [**15**]. Some of these are similar to GRR in that they allow a unique regularization parameter for each coefficient, although the penalization is in an ℓ_1-sense. Examples include the adaptive lasso for $p < n$ problems [**16**] and for the diverging parameters problem, the recent extension by [**8**]. As well, [**18**] recently introduced the adaptive elastic net, which imposes both an adaptive lasso and a ridge penalty. Similar to the original elastic net [**17**], the additional ridge penalty encourages a grouping effect that can help select groups of correlated variables and stabilizes model predictions.

Unlike lasso-based estimators, GRR imposes parameterwise adaptation using ℓ_2-regularization, which may have benefits in high-dimensional correlated settings. In the classic setting when $n > p$, it is well known that the ridge estimator uniformly shrinks the OLS, thus reducing its squared-length relative to the OLS [**6**]. This variance reduction enables the ridge estimator to outperform the OLS in correlated settings. However, the ridge estimator shrinks all coefficients uniformly to zero, which is effective only when all coefficients are small. On the other hand, GRR generalizes ridge estimation by using a unique parameter λ_k for each coefficient β_k, which allows GRR to achieve non-uniform shrinkage, thus making it feasible to selectively shrink coefficients to zero, similar to the lasso. In the classic settiing $n > p$, it has been shown that GRR estimators can achieve oracle properties [**10**].

We take a geometric approach to study the properties of GRR (Section 2). Because OLS is not uniquely defined when $p > n$, these arguments make use of the *minimum least squares* (MLS) estimator, which is the uniquely defined least squares estimator with minimum squared-length (see Definition 2.3). Using a modified MLS estimator, a novel geometric interpretation for GRR is described that lends insight into its properties. Section 3 lists implications of these findings for GRR in high

dimensions. Analogous to ridge estimation in $n > p$ settings, it is shown that the GRR estimator is shrunk relative to MLS in correlated settings and that the GRR predictor has the tendency to shrink towards zero in unfavorable directions in the $\mathbf{X}$-column space. This can lead to both improved estimation and prediction over MLS. However, unlike the classic setting, GRR is constrained to lie in a low-dimensional subspace containing the modified MLS estimator. This implies that for accurate estimation the true parameter vector should be sparse. In non-sparse situations, accurate estimation cannot be guaranteed. Section 4 summarizes our findings and presents an empirical example.

2. Geometry and properties of the GRR estimator when $p \geq n$

Here we present a novel geometric interpretation of the GRR estimator when $p \geq n$ and list some of its key properties. The following lemma plays a key role.

LEMMA 2.1. *Let $p \geq n$ and let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{X}$ where $\mathbf{U}(n \times n)$ and $\mathbf{V}(p \times n)$ are column orthonormal matrices ($\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$) and $\mathbf{D}(n \times n)$ is a diagonal matrix with entries $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$ (the singular values of $\mathbf{X}$). Let $\mathbf{A} = \mathbf{V}^T(\mathbf{Q} + \lambda\mathbf{I}_p)$. Then for any $\lambda > 0$*

$$\mathbf{A}^+ = \mathbf{V}\mathbf{S}_\lambda^{-1}, \tag{2.1}$$

where $\mathbf{S}_\lambda = \operatorname{diag}\{d_i^2 + \lambda\}_{i=1}^n$ is an $n \times n$ diagonal matrix and $\mathbf{A}^+$ denotes the Moore–Penrose [**13**] *generalized inverse of $\mathbf{A}$. Furthermore, $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_n$.*

Lemma 2.1 will enable us to re-express $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ in a manner more conducive to studying its properties. First we observe that $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ can be recast as a rescaled ridge estimator. Let $\mathbf{X}_* = \mathbf{X}\boldsymbol{\Lambda}^{-1/2}$ and $\mathbf{Q}_* = \mathbf{X}_*^T\mathbf{X}_*$ (hereafter "$*$" will be used to indicate a term mapped under the transformation $\mathbf{X}\boldsymbol{\Lambda}^{-1/2}$). Then

$$\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \boldsymbol{\Lambda}^{-1/2}(\mathbf{Q}_* + \mathbf{I}_p)^{-1}\mathbf{X}_*^T\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}\widehat{\boldsymbol{\beta}}_{\mathrm{R}}^*, \tag{2.2}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}^* = (\mathbf{Q}_* + \mathbf{I}_p)^{-1}\mathbf{X}_*^T\mathbf{Y}$ is the ridge estimator for the design matrix $\mathbf{X}_*$ with ridge parameter $\lambda = 1$. Let $\mathbf{X}_* = \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T$ be the SVD for $\mathbf{X}_*$. Let $d_{1,*} \geq \cdots \geq d_{n,*} \geq 0$ denote the diagonal elements of $\mathbf{D}_*$. Lemma 2.1 implies the following result.

THEOREM 2.2. *If $p \geq n$ and $\lambda_k > 0$ for $k = 1, \ldots, p$, then*

$$\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \boldsymbol{\Lambda}^{-1/2}\mathbf{V}_*\mathbf{S}_{*1}^{-1}\mathbf{R}_*^T\mathbf{Y}, \tag{2.3}$$

where $S_{*1} = \operatorname{diag}\{d_{i,*}^2 + 1\}_{i=1}^n$ and $\mathbf{R}_* = \mathbf{U}_*\mathbf{D}_*$. Moreover, (2.3) *can be calculated in $O(pn^2)$ operations.*

2.1. Geometry. We now describe a novel geometric interpretation for $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$. The MLS estimator will play a key role in this description and is formally defined below. For this, and all other results, we hereafter assume that $p \geq n$, $\lambda_k > 0$ for $k = 1, \ldots, p$, and $\lambda > 0$, unless otherwise stated.

DEFINITION 2.3. Call any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ a least squares solution if $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \leq \|\mathbf{Y} - \mathbf{X}\mathbf{z}\|^2$ for all $\mathbf{z} \in \mathbb{R}^p$. A vector $\boldsymbol{\beta}$ is called a MLS solution if $\boldsymbol{\beta}$ is a least squares solution and $\|\boldsymbol{\beta}\|^2 < \|\mathbf{z}\|^2$ for all other least squares solutions $\mathbf{z}$.

A celebrated result, due to [**14**], is that the MLS estimator exists and is the unique estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}} = \mathbf{X}^{+}\mathbf{Y} = \lim_{\lambda \to 0} \widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{V}\mathbf{S}_0^{+}\mathbf{R}^T\mathbf{Y},$$

where $\mathbf{R} = \mathbf{U}\mathbf{D}$ and $\mathbf{S}_0^{+} = \mathrm{diag}\{s_{0i}^{+}\}_{i=1}^{n}$ is the Moore–Penrose generalized inverse of $\mathbf{S}_0$ defined by

$$s_{0i}^{+} = \begin{cases} 1/d_i^2 & \text{if } d_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Our geometric result uses a slightly modified MLS estimator obtained using the transformed design matrix $\mathbf{X}_*$. The modified MLS estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*} = \boldsymbol{\Lambda}^{-1/2}\mathbf{X}_*^{+}\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}\mathbf{V}_*\mathbf{S}_{*0}^{+}\mathbf{R}_*^T\mathbf{Y}. \tag{2.4}$$

Note that in the special case $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p$ we obtain $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*} = \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}$. Consider the following geometric interpretation for GRR.

THEOREM 2.4. *$\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ is the solution to the following optimization problem*:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \ \mathbb{Q}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}) \text{ subject to } \boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta} \le L, \tag{2.5}$$

for some $L > 0$, where

$$\mathbb{Q}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}) = \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}\big)^T \big(\boldsymbol{\Lambda}^{1/2}\mathbf{Q}_*\boldsymbol{\Lambda}^{1/2}\big)\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}\big)$$

defines an ellipsoid with contours $\boldsymbol{\Sigma}(c) = \{\boldsymbol{\beta} : \mathbb{Q}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}) = c^2\}$ centered at $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*}$.

Theorem 2.4 shows that the GRR estimator is the solution to a constrained optimization problem involving the contours of an ellipsoid centered at the modified MLS estimator. This generalizes the classic setting $n > p$ from an optimization problem involving OLS to one involving MLS. The constraint region is generalized as well. For GRR, the constraint region is an ellipsoid that depends upon $\boldsymbol{\Lambda}$, whereas in the classic setting the constraint region is spherical. Another key difference is that the dimension of the subspace that $\widehat{\boldsymbol{\beta}}_G$ lies in depends upon n, and not p (see Theorem 2.5 below). Figure 1 provides an illustration of Theorem 2.4.

2.2. Properties. The following theorem summarizes key properties of GRR. It also gives an explicit representation for the linear predictor $\widehat{\boldsymbol{\mu}}_{\mathrm{G}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$. In the following, let $\mathbf{v}_{i,*}$ be the ith column vector of $\mathbf{V}_*$ and $\mathbf{u}_{i,*}$ be the ith column of $\mathbf{U}_*$.

THEOREM 2.5. *$\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \boldsymbol{\Lambda}^{-1/2}\sum_{i=1}^{\mathfrak{d}} d_{i,*}\eta_{i,*}\mathbf{v}_{i,*}$, where $\eta_{i,*} = (d_{i,*}^2 + 1)^{-1}\mathbf{u}_{i,*}^T\mathbf{Y}$. That is, $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ lies in the $\mathfrak{d}$-dimensional subspace $\boldsymbol{\Lambda}^{-1/2}(\mathscr{V}_*) = \{\boldsymbol{\Lambda}^{-1/2}\mathbf{v} : \mathbf{v} \in \mathscr{V}_*\}$, where $\mathfrak{d} = \mathrm{rank}(\mathbf{X}) \le n$ and $\mathscr{V}_*$ is the span of $\{\mathbf{v}_{1,*}, \ldots, \mathbf{v}_{\mathfrak{d},*}\}$; i.e., $\mathscr{V}_*$ is the span of the eigenvectors of $\mathbf{Q}_*$. The linear predictor is expressible as $\widehat{\boldsymbol{\mu}}_{\mathrm{G}} = \sum_{i=1}^{\mathfrak{d}} d_{i,*}^2\eta_{i,*}\mathbf{u}_{i,*}$.*

REMARK 2.6. Using (2.4) and similar arguments as in the proof of Theorem 2.5 the modified MLS is expressible as $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^{*} = \boldsymbol{\Lambda}^{-1/2}\sum_{i=1}^{\mathfrak{d}} d_{i,*}^{-1}(\mathbf{u}_{i,*}^T\mathbf{Y})\mathbf{v}_{i,*}$ and its linear predictor can be written as $\widehat{\boldsymbol{\mu}}_{\mathrm{MLS}}^{*} = \sum_{i=1}^{\mathfrak{d}}(\mathbf{u}_{i,*}^T\mathbf{Y})\mathbf{u}_{i,*}$. These facts will become handy later in Section 3.

3. Implications for GRR when $p \ge n$

We now list several interesting facts that follow from our previous results.

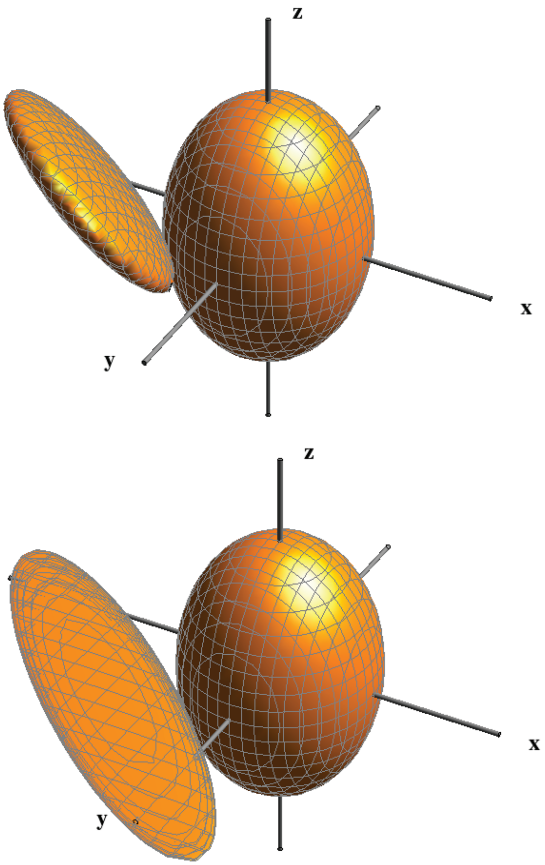

FIGURE 1. Illustration of GRR geometry. Top figure corresponds to a simulation where $p = 100$, $n = 25$, and $\beta_{0,k} = 0$ for $k \geq 3$ and $\lambda_k = \infty$ for $k \geq 4$. Only the first 3 coordinates of $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ are nonzero, and these equal the point where the ellipsoid first touches the elliptical constraint region centered at zero. Bottom figure is $\lambda_k = \infty$ for $k \geq 3$. Now only the first two coordinates of $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ are nonzero (the point touching the elliptical constraint region constrained to the x, y-plane and centered at zero).

3.1. Efficient calculation. In itself, Theorem 2.2 has immediate practical value as it permits efficient calculation of GRR in a linear number of operations in p; thus making GRR computationally feasible in $p \gg n$ settings. Note that as a special case of (2.3) the following representation for the ridge estimator holds

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{V}\mathbf{S}_{\lambda}^{-1}\mathbf{R}^{T}\mathbf{Y}. \tag{3.1}$$

Similar to (2.3), this shows that the ridge estimator can be computed in a linear number of operations. See [**4**] for related work.

3.2. Shrinkage. The expression for $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$ in Theorem 2.5 shows that the GRR predictor applies the greatest amount of shrinkage to those columns of $\mathbf{U}_*$ with smallest singular values. Thus, GRR is attempting to shrink the predictor in unfavorable directions relative to the column space of $\mathbf{X}_*$. This is a generalization of a well-known property of ridge regression (see, for example, [**5**, Chapter 3]). We can demonstrate the effect of this shrinkage by comparing the GRR predictor to the modified MLS predictor (we could compare GRR to MLS, but this is not as straightforward, and this extra effort may be unnecessary as the modified MLS estimator has been reported to have similar empirical behavior to the MLS; see [**12**]). By Remark 2.6 and Theorem 2.5, we have

$$\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}} = \sum_{i=1}^{\mathfrak{d}} (\mathbf{u}_{i,*}^T \mathbf{Y})\, \mathbf{u}_{i,*} \tag{3.2}$$

$$\widehat{\boldsymbol{\mu}}_{\mathrm{G}} = \sum_{i=1}^{\mathfrak{d}} \frac{\delta_{i,*}}{\delta_{i,*}+1} (\mathbf{u}_{i,*}^T \mathbf{Y})\, \mathbf{u}_{i,*}, \tag{3.3}$$

where $\delta_{i,*} = d_{i,*}^2$. Notice how $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$ is shrunk in $\mathbf{u}_{i,*}$-directions corresponding to small singular values. Expressions (3.2) and (3.3) also show that the length of $\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}}$ is always larger than $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$. Indeed, taking expectations, $\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}}$ is always larger on average, because

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}}\|^2 = \mathbb{E}\|\widehat{\boldsymbol{\mu}}_{\mathrm{G}}\|^2 + \sum_{i=1}^{\mathfrak{d}} \frac{2\delta_{i,*}+1}{(\delta_{i,*}+1)^2}(\mu_{i,*}^2 + \sigma_0^2),$$

where $\mu_{i,*} = \mathbf{u}_{i,*}^T \boldsymbol{\mu}$.

Theorem 2.5 also generalizes the well known property of GRR as a shrinkage estimator. By Remark 2.6, the modified MLS estimator is expressible as

$$\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}} = \boldsymbol{\Lambda}^{-1/2} \sum_{i=1}^{\mathfrak{d}} (d_{i,*} + d_{i,*}^{-1}) \eta_{i,*} \mathbf{v}_{i,*}. \tag{3.4}$$

Comparing this to $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$, we see that each term in the summation of $\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}$ is larger than that of $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ by an amount $d_{i,*}^{-1}$. Multiplying $\boldsymbol{\Lambda}^{1/2}$ throughout both expressions,

$$\|\boldsymbol{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}\|^2 = \|\boldsymbol{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}\|^2 + \sum_{i=1}^{\mathfrak{d}} (2 + \delta_{i,*}^{-1})\eta_{i,*}^2.$$

Thus, under a $\boldsymbol{\Lambda}^{1/2}$ rescaling, the squared-length of $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ is always smaller than $\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}$. In particular, $\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}$ becomes elongated in the presence of small singular values. This is a generalization of the classic $n > p$ setting. There, it is well known that the ridge regression uniformly shrinks the OLS toward zero, so that its squared-length is always smaller than that of the OLS [**6**].

3.3. Sparsity. Theorem 2.5 shows that $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ lies within the subspace $\boldsymbol{\Lambda}^{-1/2}(\mathscr{V}_*)$ containing the modified MLS estimator: this forces $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ to lie in a low $\mathfrak{d}$-dimensional subspace which may degrade its performance in high dimensional non-sparse settings. To see why, consider the distance between the scaled GRR estimator, $\boldsymbol{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$, and the scaled regression parameter, $\boldsymbol{\beta}_0^* = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\beta}_0$. We decompose $\boldsymbol{\beta}_0^*$ into its projection onto $\mathscr{V}_*$ and the orthogonal subspace $\mathscr{V}_*^{\perp}$. Because $\boldsymbol{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}_{\mathrm{G}} \in \mathscr{V}_*$,

the distance between $\mathbf{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ and $\boldsymbol{\beta}_0^*$ can be bounded below by the projection of $\boldsymbol{\beta}_0^*$ onto $\mathscr{V}_*^{\perp}$. Consequently

$$\left\|\mathbf{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{G}}-\boldsymbol{\beta}_0)\right\|_2^2 \geq \left\|\left(\mathbf{I}_p-\mathbf{V}_*\mathbf{D}_*(\mathbf{V}_*\mathbf{D}_*)^{+}\right)\boldsymbol{\beta}_0^*\right\|^2.$$

If $\boldsymbol{\beta}_0$ lies in a high-dimensional subspace, then it may not be possible to find a $\mathbf{\Lambda}$ to make this distance zero. On the other hand, if $\boldsymbol{\beta}_0$ sits in a low-dimensional subspace of dimension no larger than $\mathfrak{d}$, then there always exists a ridge matrix making the right-hand side zero. The dimensionality of $\boldsymbol{\beta}_0$ is a sparsity condition. Because $\mathfrak{d} \leq n$, accurate estimation is guaranteed only when the sparsity condition $p_0 \leq n$ is met, where p_0 equals the number of nonzero coefficients of $\boldsymbol{\beta}_0$. But in non-sparse conditions, where $p_0 > n$, no such guarantee holds.

3.4. Prediction. It is reasonable to expect that GRR will outperform traditional least squares (i.e., MLS) in high dimensions. To formally investigate this we consider the difference in prediction performance of the GRR to the modified MLS (as we have remarked, working directly with the MLS is difficult, thus we instead work with the modified MLS which serves as a reasonable proxy). In the following, let $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}\boldsymbol{\beta}_0)$ be the true predicted value.

THEOREM 3.1.

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}}_{\mathrm{MLS}}^*-\boldsymbol{\mu}\|^2 = \mathbb{E}\|\widehat{\boldsymbol{\mu}}_{\mathrm{G}}-\boldsymbol{\mu}\|^2 + \sum_{i=1}^{\mathfrak{d}} \frac{(2\delta_{i,*}+1)\sigma_0^2-\mu_{i,*}^2}{(\delta_{i,*}+1)^2}. \tag{3.5}$$

Theorem 3.1 gives the mean-squared error for the modified MLS relative to GRR and hence provides insight into the risk behavior for the MLS relative to GRR. Interestingly, (3.5) identifies scenarios where the risk for $\widehat{\boldsymbol{\mu}}_{\mathrm{MLS}}^*$ may be smaller than $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$. One example is a noiseless system in which $\sigma_0^2 = 0$. Then the second term on the right of (3.5) is negative and $\widehat{\boldsymbol{\mu}}_{\mathrm{MLS}}^*$ will have smaller risk than $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$. In general, however, the risk for $\widehat{\boldsymbol{\mu}}_{\mathrm{MLS}}^*$ will be larger than $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$ if σ_0^2 is nonzero and $\delta_{i,*}$ is large — the latter occurs if the singular values are large. Thus, outside of low noise, high signal systems, the risk for $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$ is expected to be smaller.

3.5. Estimation. In a similar fashion we can compare the MSE performance of GRR to the modified MLS estimator. Consider the following MSE decomposition.

THEOREM 3.2. *Let* $\alpha_i = (\mathbf{\Lambda}^{1/2}\boldsymbol{\beta}_0)^T\mathbf{v}_{i,*}$. *Then*

$$\begin{aligned}\mathbb{E}\|\mathbf{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^*-\boldsymbol{\beta}_0)\|^2 \\ = \mathbb{E}\|\mathbf{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{G}}-\boldsymbol{\beta}_0)\|^2 + \sum_{i=1}^{\mathfrak{d}} \frac{(2\delta_{i,*}+1)(\mu_{i,*}^2+\sigma_0^2)-2\alpha_i\delta_{i,*}^{1/2}(\delta_{i,*}+1)\mu_{i,*}}{\delta_{i,*}(\delta_{i,*}+1)^2}.\end{aligned} \tag{3.6}$$

Interpreting (3.6) is not as straightforward as (3.5). However, one interesting conclusion is that unlike the prediction scenario, a noiseless system with $\sigma_0^2 = 0$ does not necessarily confer a MSE advantage for $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^*$. Also, from the second term in (3.6), we see that small singular values will inflate the MSE for $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^*$ and that this inflation is further enhanced by the presence of $\mu_{i,*}^2$ (and the term σ_0^2). These results are consistent with our earlier comments that small singular values create instability in $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^*$. Thus we expect $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}^*$ (and consequently $\widehat{\boldsymbol{\beta}}_{\mathrm{MLS}}$) to have poor MSE performance in high-dimensional correlated scenarios.

4. Discussion

The p bigger than n setting presents an ill-conditioned scenario where spurious correlations between variables can make estimation and prediction difficult. This paper studied how GRR estimators would perform in this setting given their stabilizing effects seen in $n > p$ situations.

Using geometric arguments, it was shown that the properties of GRR when $p > n$ shared similar features to the solution in the classic $n > p$ setting but also differed in several important ways. Like the classic setting, shrinkage plays a role which can lead to both improved estimation and prediction over MLS (least squares). However, an important difference in high dimensions is that the GRR solution is constrained to lie in a subspace containing the MLS estimator of dimension at most n (as opposed to a subspace of dimension p in the classic setting). This implies that for accurate estimation the true parameter vector should be sparse in the sense that $p_0 \leq n$. In non-sparse situations, accurate estimation cannot be guaranteed.

The high-dimensional sparse setting has attracted a considerable amount of research interest with a large focus on lasso and lasso-type regularization. Our results suggest that GRR can also have excellent performance in such settings if the ridge matrix is selected appropriately. One way to proceed would be to use a Bayesian approach which naturally lend themselves to ridge estimation. In particular, let $\mathbf{\Gamma} = \text{diag}\{\gamma_k\}_{k=1}^p$ be a $p \times p$ diagonal matrix with diagonal entries $\gamma_k > 0$. Consider the following Bayesian normal-normal hierarchy:

$$\begin{aligned} (\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) &\sim \text{N}(X\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ (\boldsymbol{\beta} \mid \mathbf{\Gamma}) &\sim \text{N}(\mathbf{0}, \mathbf{\Gamma}). \end{aligned} \tag{4.1}$$

Conjugacy ensures that the posterior distribution for $\boldsymbol{\beta}$ is multivariate normal. By standard calculations, the posterior distribution of $\boldsymbol{\beta}$ is

$$(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{\Gamma}, \sigma^2) \sim \text{N}(\boldsymbol{\mu}_{\mathbf{\Gamma}}, \sigma^2 \mathbf{\Sigma}_{\mathbf{\Gamma}})$$

where $\boldsymbol{\mu}_{\mathbf{\Gamma}} = \mathbf{\Sigma}_{\mathbf{\Gamma}} \mathbf{X}^T \mathbf{Y}$ is the posterior mean of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}_{\mathbf{\Gamma}} = (\mathbf{Q} + \sigma^2 \mathbf{\Gamma}^{-1})^{-1}$. Observe that $\boldsymbol{\mu}_{\mathbf{\Gamma}}$ is a GRR estimator.

Model (4.1) is a "plain vanilla" Bayesian hierarchy that assumes a fixed ridge matrix. However, in practice a more sophisticated hierarchy using a prior for $\mathbf{\Gamma}$ is a preferred way to estimate $\boldsymbol{\beta}$. This reduces the risk of poor estimation if $\mathbf{\Gamma}$ is misspecified. In particular, a non-generate prior for $\mathbf{\Gamma}$ results in a posterior mean that is no longer a GRR estimator but instead is a weighted (averaged) GRR estimator (WGRR). Such estimators are closely tied to mixed GRR estimators which are known to have a minimax property [**12**]. An example of such a hierarchy is the rescaled spike and slab model used in [**11**] which utilizes a continuous bimodal prior for $\mathbf{\Gamma}$ that allows γ_k to be adaptively determined.

We illustrate this method using the diabetes data of [**2**]; a popular benchmark dataset used in regression. The data consists of $n = 442$ patients in which the response is a quantitative measure of disease progression for a patient. The original data included 10 baseline measurements for each patient, age, sex, body mass index, average blood pressure and six blood serum measurements, in addition to 45 interactions formed by all pairwise interactions of the 10 baseline variables and 9 quadratic terms for the 9 baselines measurements that were continuous. To this we added 1000 "noise" variables, each sampled independently from a multivariate

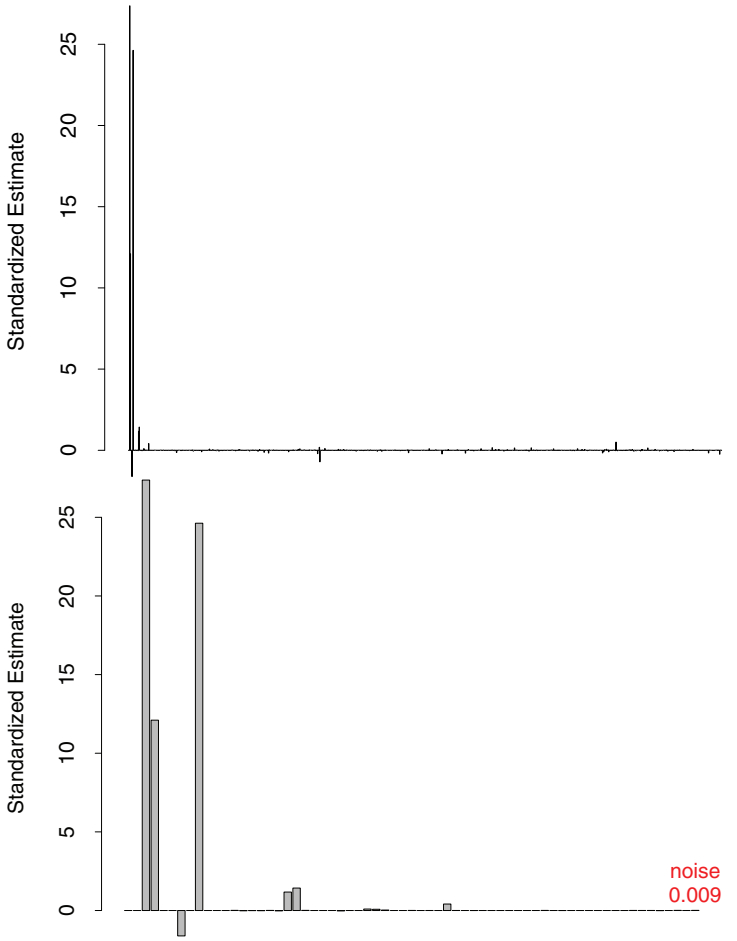

FIGURE 2. Standardized coefficient estimates for diabetes data with $q = 1000$ correlated noise variables ($n = 442$, $p = 1064$). Coefficient estimates obtained using a WGRR estimator. Top figure displays all $p = 1064$ coefficient estimates. Bottom figure displays the original 64 variables with right most estimate labeled "noise" equaling the mean absolute coefficient value of the $q = 1000$ noise variables.

normal distribution with mean zero and equicorrelation matrix with correlation $\rho = 0.5$. In total our modified data contained $p = 1064$ variables.

The standardized posterior coefficient values are displayed in Figure 2. The results were obtained using the R-package "spikeslab" [**9**] which fits a rescaled and slab model. All variables were standardized to have a sample mean of zero and sample variance of one. This yields standardized posterior coefficient values that can be compared against one another. The top figure displays all $p = 1064$ variables while the bottom figure displays the coefficient estimates for the original 64 variables. The right most estimate labeled "noise" is the averaged value of the absolute posterior coefficient estimates for the 1000 noise variables. Even in the presence

of high correlation, the posterior estimates are shrunk towards zero for nearly all noise variables (average of.009) and only a subset of the original 64 variables appear informative. For technical details and further empirical illustrations the interested reader should consult [**11**].

Appendix A. Proofs

PROOF OF LEMMA 2.1. We first show that $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_n$. Using $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$, $\mathbf{Q} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, and $\mathbf{V}\mathbf{S}_\lambda^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_n)^{-1}$, deduce that

$$\begin{aligned}\mathbf{A}\mathbf{A}^+ &= (\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{V}^T)\big[\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_n)^{-1}\big] \\ &= (\mathbf{D}^2 + \lambda\mathbf{I}_n)(\mathbf{D}^2 + \lambda\mathbf{I}_n)^{-1} \\ &= \mathbf{I}_n.\end{aligned}$$

From this it immediately follows that: (i) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$; (ii) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$; and (iii) $(\mathbf{A}\mathbf{A}^+)^T = \mathbf{A}\mathbf{A}^+$. Furthermore,

$$\mathbf{A}^+\mathbf{A} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_n)^{-1}(\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{V}^T) = \mathbf{V}\mathbf{V}^T.$$

Therefore: (iv) $(\mathbf{A}^+\mathbf{A})^T = \mathbf{V}\mathbf{V}^T = \mathbf{A}^+\mathbf{A}$. Properties (i)–(iv) are the four criteria required for a Moore–Penrose generalized inverse. □

PROOF OF THEOREM 2.2. By (2.2), the GRR estimator can be expressed as a ridge estimator scaled by a diagonal matrix. Therefore, it suffices to prove that (3.1) holds and that the number of operations required to calculate (3.1) is of order $O(pn^2)$. Set $\mathbf{\Lambda} = \lambda\mathbf{I}_p$. Taking the derivative with respect to $\boldsymbol{\beta}$ in (1.2), setting this to zero, and multiplying right and left-hand sides by $\mathbf{V}^T$, it follows that $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$ must satisfy

$$\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{V}^T\mathbf{X}^T\mathbf{Y}.$$

The solution must be $\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{A}^+\mathbf{V}^T\mathbf{X}^T\mathbf{Y}$ because upon substituting this into the left-hand side we obtain

$$\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{A}\mathbf{A}^+\mathbf{V}^T\mathbf{X}^T\mathbf{Y} = \mathbf{V}^T\mathbf{X}^T\mathbf{Y},$$

where we have used the fact that $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_n$ from Lemma 2.1. Now substituting the right-hand side of (2.1) for $\mathbf{A}^+$, yields

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \mathbf{V}\mathbf{S}_\lambda^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{Y} = \mathbf{V}\mathbf{S}_\lambda^{-1}\mathbf{R}^T\mathbf{Y}.$$

To determine the number of operations required to compute $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$, note that the SVD for $\mathbf{X}$ requires $O(pn^2)$ operations. Once the SVD is obtained, inverting $\mathbf{S}_\lambda^{-1}$ requires $O(n)$ operations. Multiplying this (of size $n \times n$) by $\mathbf{V}(p \times n)$ requires $O(pn^2)$ operations (note that because S_λ^{-1} is diagonal, this can be reduced further to $O(pn)$ operations, but this level of refinement is not essential). Multiplying by $\mathbf{R}^T\mathbf{Y}$ requires a total of $O(pn)$ operations. The total number of operations equals

$$O(pn^2) + O(n) + O(pn^2) + O(pn) = O(pn^2).$$ □

PROOF OF THEOREM 2.4. By (2.2), $\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \mathbf{\Lambda}^{-1/2}\widehat{\boldsymbol{\beta}}^*_{\mathrm{R}}$ where $\widehat{\boldsymbol{\beta}}^*_{\mathrm{R}}$ is the ridge estimator under X_* with ridge parameter $\lambda = 1$. We show that $\mathbf{\Lambda}^{-1/2}\widehat{\boldsymbol{\beta}}^*_{\mathrm{R}}$ is the solution to (2.5). As a Lagrangian problem, (2.5) can be written as

$$\underset{(\boldsymbol{\beta},\ell)\in\mathbb{R}^p\times\mathbb{R}_+}{\text{minimize}} \big\{\mathbb{Q}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}) + \ell(\boldsymbol{\beta}^T\mathbf{\Lambda}\boldsymbol{\beta} - L)\big\},$$

where ℓ is the Lagrangian multiplier. Because L is arbitrary we can assume that $\ell = 1$ without loss of generality. Taking the derivative with respect to $\boldsymbol{\beta}$, the solution is

$$2\boldsymbol{\Lambda}\boldsymbol{\beta} + 2\boldsymbol{\Lambda}^{1/2}\mathbf{Q}_*\boldsymbol{\Lambda}^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}}) = \mathbf{0}.$$

Multiplying throughout by $\boldsymbol{\Lambda}^{-1/2}$, $\boldsymbol{\beta}$ must satisfy

$$\begin{aligned}(\mathbf{Q}_* + \mathbf{I}_p)\boldsymbol{\Lambda}^{1/2}\boldsymbol{\beta} &= \mathbf{Q}_*\boldsymbol{\Lambda}^{1/2}\widehat{\boldsymbol{\beta}}^*_{\mathrm{MLS}} \\ &= \mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T\mathbf{V}_*\mathbf{S}_{*0}^+\mathbf{R}_*^T\mathbf{Y} \\ &= \mathbf{V}_*\mathbf{D}_*^2\mathbf{S}_{*0}^+\mathbf{D}_*\mathbf{U}_*^T\mathbf{Y} \\ &= \mathbf{V}_*\mathbf{D}_*\mathbf{U}_*^T\mathbf{Y} \\ &= \mathbf{X}_*^T\mathbf{Y}.\end{aligned}$$

Therefore, $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-1/2}\widehat{\boldsymbol{\beta}}^*_{\mathrm{R}}$; thus verifying that (2.5) is the optimization problem for $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$. □

PROOF OF THEOREM 2.5. By (2.3), we can write $\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \boldsymbol{\Lambda}^{-1/2}\mathbf{V}_*\mathbf{S}_{*1}^{-1}\mathbf{D}_*\mathbf{U}_*^T\mathbf{Y}$. The stated representation for $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ follows with some simple rearrangement. From this it is clear that $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ lies in the subspace $\mathscr{V}_*$ (and hence $\boldsymbol{\Lambda}^{-1/2}(\mathscr{V}_*)$). To see that $\mathscr{V}_*$ can be interpreted as the span of the eigenvectors of $\mathbf{Q}_*$, note that $\mathbf{Q}_* = \mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T$. Thus, $\mathbf{Q}_*\mathbf{v}_{i,*} = d_{i,*}^2\mathbf{v}_{i,*}$, and hence $\mathbf{v}_{i,*}$ is an eigenvector of Q_* under the condition that $d_{i,*}^2 > 0$. Finally, to prove the claim for $\widehat{\boldsymbol{\mu}}_{\mathrm{G}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$, using $\mathbf{X}_* = \mathbf{X}\boldsymbol{\Lambda}^{-1/2}$ and the representation for $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$, we have

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_{\mathrm{G}} &= \mathbf{X}^*\mathbf{V}_*\mathbf{S}_{*1}^{-1}\mathbf{D}_*\mathbf{U}_*^T\mathbf{Y} \\ &= \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T\mathbf{V}_*\mathbf{S}_{*1}^{-1}\mathbf{D}_*\mathbf{U}_*^T\mathbf{Y} \\ &= \mathbf{U}_*(\mathbf{D}_*\mathbf{S}_{*1}^{-1}\mathbf{D}_*)\mathbf{U}_*^T\mathbf{Y}.\end{aligned}$$

It is easily checked that this corresponds to the stated expression for $\widehat{\boldsymbol{\mu}}_{\mathrm{G}}$. □

PROOF OF THEOREM 3.1. We can write $\boldsymbol{\mu}$ in terms of the orthonormal basis $(\mathbf{u}_{i,*})_{i=1}^n$. We have $\boldsymbol{\mu} = \sum_{i=1}^n \mu_{i,*}\mathbf{u}_{i,*}$. Therefore, by (3.2) and (3.3),

$$\begin{aligned}\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}} - \boldsymbol{\mu} &= \sum_{i=1}^{\mathfrak{d}}\left(\frac{\delta_{i,*}}{\delta_{i,*}+1} - \frac{\mu_{i,*}}{\mathbf{u}_{i,*}^T\mathbf{Y}}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{u}_{i,*} \\ &\qquad + \sum_{i=1}^{\mathfrak{d}}\left(1 - \frac{\delta_{i,*}}{\delta_{i,*}+1}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{u}_{i,*} \\ &= (\widehat{\boldsymbol{\mu}}_{\mathrm{G}} - \boldsymbol{\mu}) + \sum_{i=1}^{\mathfrak{d}}\left(\frac{1}{\delta_{i,*}+1}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{u}_{i,*}.\end{aligned}$$

Squaring and collecting terms, deduce that

$$\begin{aligned}&\|\widehat{\boldsymbol{\mu}}^*_{\mathrm{MLS}} - \boldsymbol{\mu}\|^2 \\ &\qquad = \|\widehat{\boldsymbol{\mu}}_{\mathrm{G}} - \boldsymbol{\mu}\|^2 + \sum_{i=1}^{\mathfrak{d}}\frac{(2\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})^2 - 2\mu_{i,*}(\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})}{(\delta_{i,*}+1)^2}.\end{aligned} \tag{A.1}$$

One can easily verify that

$$\begin{aligned}(\text{A.2})\quad \mathbb{E}[(2\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})^2 - 2\mu_{i,*}(\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})] \\ = (2\delta_{i,*}+1)(\mu_{i,*}^2+\sigma_0^2) - 2\mu_{i,*}(\delta_{i,*}+1)\mu_{i,*} \\ = (2\delta_{i,*}+1)\sigma_0^2 - \mu_{i,*}^2.\end{aligned}$$

The theorem is proved by using this, and taking expecations in (A.1). □

PROOF OF THEOREM 3.2. The proof is similar to that of Theorem 3.1. Note that $(\mathbf{v}_{i,*})_{i=1}^{\mathfrak{d}}$ is an orthonormal basis for $\mathscr{V}_*$. Let $\boldsymbol{\beta}_0^* = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\beta}_0$ and write $\mathscr{V}_*^\perp$ for the orthogonal subspace to $\mathscr{V}_*$. Then, $\boldsymbol{\beta}_0^* = \sum_{i=1}^n \alpha_i \mathbf{v}_{i,*} + \Delta^*$, where Δ^* is the projection of $\boldsymbol{\beta}_0^*$ onto $\mathscr{V}_*^\perp$. Using the representation (3.4) for $\widehat{\boldsymbol{\beta}}_{\text{MLS}}^*$ and the representation for $\widehat{\boldsymbol{\beta}}_{\text{G}}$ given in Theorem 2.5, we have

$$\begin{aligned}\boldsymbol{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{MLS}}^* - \boldsymbol{\beta}_0) &= \sum_{i=1}^{\mathfrak{d}}\left(\frac{d_{i,*}}{\delta_{i,*}+1} - \frac{\alpha_i}{\mathbf{u}_{i,*}^T\mathbf{Y}}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{v}_{i,*} - \Delta^* \\ &\qquad + \sum_{i=1}^{\mathfrak{d}}\left(\frac{d_{i,*}+d_{i,*}^{-1}}{\delta_{i,*}+1} - \frac{d_{i,*}}{\delta_{i,*}+1}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{v}_{i,*} \\ &= \boldsymbol{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{G}} - \boldsymbol{\beta}_0) + \sum_{i=1}^{\mathfrak{d}}\left(\frac{d_{i,*}^{-1}}{\delta_{i,*}+1}\right)(\mathbf{u}_{i,*}^T\mathbf{Y})\,\mathbf{v}_{i,*}.\end{aligned}$$

Squaring, collecting terms, and taking expectations, deduce that

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{MLS}}^* - \boldsymbol{\beta}_0)\|^2 \\ &\quad = \mathbb{E}\|\boldsymbol{\Lambda}^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{G}} - \boldsymbol{\beta}_0)\|^2 + \mathbb{E}\left[\sum_{i=1}^{\mathfrak{d}}\frac{(2\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})^2 - 2\alpha_i d_{i,*}(\delta_{i,*}+1)(\mathbf{u}_{i,*}^T\mathbf{Y})}{\delta_{i,*}(\delta_{i,*}+1)^2}\right].\end{aligned}$$

The result follows by taking the expectation inside the sum and using $\mathbb{E}(\mathbf{u}_{i,*}^T\mathbf{Y})^2 = \mu_{i,*}^2 + \sigma_0^2$ and $\mathbb{E}(\mathbf{u}_{i,*}^T\mathbf{Y}) = \mu_{i,*}$. □

References

[1] T. T. Cai and J. Lv, *Discussion: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist.* **35** *(2007), no. 6, 2313–2351; MR2382644] by E. Candes and T. Tao*, Ann. Statist. **35** (2007), no. 6, 2365–2369, DOI 10.1214/009053607000000442. MR2382647 (2009b:62015)

[2] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression. With discussion, and a rejoinder by the authors*, Ann. Statist. **32** (2004), no. 2, 407–499, DOI 10.1214/009053604000000067. MR2060166 (2005d:62116)

[3] J. Fan and J. Lv, *Sure independence screening for ultrahigh dimensional feature space*, J. R. Stat. Soc. Ser. B Stat. Methodol. **70** (2008), no. 5, 849–911, DOI 10.1111/j.1467-9868.2008.00674.x. MR2530322

[4] T. Hastie and R. Tibshirani, *Efficient quadratic regularization for expression arrays*, Biostatistics **5** (2004), no. 3, 329–340, DOI 10.1093/biostatistics/kxh010.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning. Data mining, inference, and prediction*, Springer Series in Statistics, Springer-Verlag, New York, 2001. MR1851606 (2002k:62048)

[6] A. E. Hoerl and R. W. Kennard, *Ridge regression*: *Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67, DOI 10.1080/00401706.1970.10488634.

[7] A. E. Hoerl and R. W. Kennard, *Ridge regression*: *Applications to nonorthogonal problems.*, Technometrics **12** (1970), no. 1, 69–82; Erratum, Technometrics **12**, no. 3, 723.

[8] J. Huang, S. Ma, and C.-H. Zhang, *Adaptive Lasso for sparse high-dimensional regression models*, Statist. Sinica **18** (2008), no. 4, 1603–1618. MR2469326 (2010a:62214)
[9] H. Ishwaran, U. B. Kogalur, and S. J. Rao, *spikeslab*: *Prediction and variable selection using spike and slab regression*, R Journal **7** (2010), no. 2, 68–73.
[10] H. Ishwaran and J. Sunil Rao, *Consistency of spike and slab regression*, Statist. Probab. Lett. **81** (2011), no. 12, 1920–1928, DOI 10.1016/j.spl.2011.08.005. MR2845909 (2012h:62258)
[11] H. Ishwaran and J. S. Rao, *Generalized ridge regression: geometry and computational solutions when p is larger than n*, Technical Report 01/2011, Division of Biostatistics, University of Miami, 2011.
[12] H. Ishwaran and J. S. Rao, *Mixing generalized ridge regressions*, unpublished.
[13] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc. **51** (1955), 406–413. MR0069793 (16,1082a)
[14] R. Penrose, *On best approximation solutions of linear matrix equations*, Proc. Cambridge Philos. Soc. **52** (1956), 17–19. MR0074092 (17,536d)
[15] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288. MR1379242 (96j:62134)
[16] H. Zou, *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc. **101** (2006), no. 476, 1418–1429, DOI 10.1198/016214506000000735. MR2279469 (2008d:62024)
[17] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol. **67** (2005), no. 2, 301–320, DOI 10.1111/j.1467-9868.2005.00503.x. MR2137327
[18] H. Zou and H. H. Zhang, *On the adaptive elastic-net with a diverging number of parameters*, Ann. Statist. **37** (2009), no. 4, 1733–1751, DOI 10.1214/08-AOS625. MR2533470 (2010j:62210)

Division of Biostatistics, Department of Public Health Sciences, University of Miami, Miami Florida 33136
E-mail address: hemant.ishwaran@gmail.com

Division of Biostatistics, Department of Public Health Sciences, University of Miami, Miami Florida 33136
E-mail address: rao.jsunil@gmail.com