Shrinkage estimation, model averaging, and degrees of freedom

Keith Knight
University of Toronto
keith@utstat.toronto.edu

Abstract

Leamer and Chamberlain (1976) showed that the ridge regression estimate is a weighted sum of ordinary least squares estimates based on subsets of predictors. In this paper, we define (a) the notion of "level of expression" of predictors in ridge regression, and (b) the notion of a "spectrum" of the ridge regression estimate and extend these notions to other linear shrinkage estimates.

1 Introduction

We consider estimation in the linear regression model

$$Y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters.

Ridge regression (Hoerl and Kennard, 1970) is often used in the case where the regression design has a high degree of multicollinearity that inflates the variance of ordinary least squares (OLS) estimation. The idea behind ridge regresson is to shrink OLS estimates towards 0, which increases their absolute bias while simultaneously reducing their variance. The shrinkage is controlled by a tuning parameter; the goal is to find a value of the tuning parameter that minimizes or nearly minimizes the mean square error (MSE). See Hoerl (2020) for an historical perspective on ridge regression.

Leamer and Chamberlain (1976) show that the ridge regression estimate can be expressed as a weighted average of OLS estimates of all (2^p) sub-models; the weights depend only on the design matrix and can be manipulated in a number of ways to quantify the importance of predictors as well as to define the weights for different model sizes. The key to all these results comes from the work of Jacobi (1841) who showed that least squares estimates can be expressed as a weighted average of so-called elemental estimates.

The form of the ridge regression estimate allows us possible to think of it as a sort of model or ensemble averaging method. Model averaging methods are often used in Bayesian inference whereby several models each having unknown parameters are combined in a hierarchical model with prior distributions on the parameters for each model as well as a prior distribution on the models themselves; see George and McCulloch (1993) and Raftery *et al.* (1997) for details. A related idea is ensemble averaging in machine learning whereby predictions from several models are combined (usually via some sort of averaging) to form a composite prediction, which hopefully will have superior statistical properties. Examples include bagging (Breiman, 1996), stacking (Wolpert, 1992), and random forests (Breiman, 2001).

The paper is organized as follows: In sections 2 and 3, we show how the model averaging result of Leamer and Chamberlain can be used to define probabilities of inclusion ("levels of expression") for the predictors as well as to give a decomposition of the "degrees of freedom" for ridge regression estimate. This decomposition is extended to more general linear estimates in section 4 and applications to twicing and boosting are considered in section 5. Some additional miscellaneous topics are considered in section 6.

2 Ridge regression as model averaging

In ridge regression, the standard practice is to normalize the predictors to have mean 0 and equal variances. The centering allows us to estimate the intercept parameter β_0 in (1) by $\bar{y}$, the average of $y_1, \cdots, y_n$; thus by subtracting $\bar{y}$ from $y_1, \cdots, y_n$, we can remove β_0 from the model. In addition, we typically scale the predictors $\{\boldsymbol{x}_i : i = 1, \cdots, n\}$ are so that the diagonal elements of the matrix $\mathcal{X}^T\mathcal{X}$ are all 1 where $\mathcal{X}$ is the $n \times p$ matrix whose i-th row is $\boldsymbol{x}_i^T$. Thus $\mathcal{X}^T\mathcal{X}$ is the correlation matrix of the p predictors. (While this standardization is useful in practice, the results given below do not depend on it.) If the rank of $\mathcal{X}$ is less than p (for example, if p is larger than n) then $\mathcal{X}^T\mathcal{X}$ is not invertible.

The ridge regression estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ is then defined as the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

where $\lambda \geq 0$ is a tuning parameter that controls the shrinkage of the OLS estimates. In matrix form, we have

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T\boldsymbol{y} \tag{2}$$

where $\boldsymbol{y} = (y_1 \ \cdots \ y_n)^T$. In the case where $\mathcal{X}^T\mathcal{X}$ is singular, we can consider the limit of $\widehat{\boldsymbol{\beta}}(\lambda)$ as $\lambda \to 0$ and the resulting estimate is the OLS estimate with minimum L_2 norm; Hastie *et al.* (2019) refer to this as "ridgeless" least squares regression. We can also allow λ to be negative (provided that $(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}$ exists) and we will briefly consider this non-standard case later.

Subsequently, it will be convenient to define $\widehat{\boldsymbol{\beta}}(\lambda)$ in terms of the augmented design matrix

$$X_\lambda = \begin{pmatrix} \mathcal{X} \\ \lambda^{1/2} I \end{pmatrix},$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ can be written as

$$\widehat{\boldsymbol{\beta}}(\lambda) = (X_\lambda^T X_\lambda)^{-1} X_\lambda^T \boldsymbol{y}^* = (\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} X_\lambda^T \boldsymbol{y}^*$$

where $\boldsymbol{y}^* = (y_1 \ \cdots \ y_n \ 0 \ \cdots \ 0)^T$. We can also define the projection matrix

$$H_\lambda = X_\lambda (\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} X_\lambda^T = \begin{pmatrix} \mathcal{X}(\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} \mathcal{X}^T & \lambda^{1/2} \mathcal{X}(\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} \\ \lambda^{1/2} (\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} \mathcal{X}^T & \lambda (\mathcal{X}^T \mathcal{X} + \lambda I)^{-1} \end{pmatrix} \tag{3}$$

The ridge regression estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ can be written as a weighted average of elemental estimates based on the rows of the matrix X_λ. Specifically, define $\mathbf{s} = \{i_1 < \cdots < i_p\}$ to be a subset of $\{1, \cdots, n, n+1, \cdots, n+p\}$ and $X_\lambda(\mathbf{s})$ to be the sub-matrix of X_λ whose rows lie in $\mathbf{s}$ and $\boldsymbol{y}^*(\mathbf{s})$ to be the sub-vector of $\boldsymbol{y}^*$ whose indices lie in $\mathbf{s}$. We can then define

$$\widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s}) = X_\lambda^{-1}(\mathbf{s}) \boldsymbol{y}^*(\mathbf{s})$$

if the inverse exists. Note that when the subset $\mathbf{s}$ contains elements $n + j_1, \cdots, n + j_k$ then the $j_1, \cdots, j_k$ components of $\widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s})$ will be exactly 0; in this case, the estimate $\widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s})$ is effectively a subset estimate based on $p - k$ predictors.

From Jacobi (1841) (see also Subrahmanyan (1972) and Mayo and Gray (1997)), we can write

$$\widehat{\boldsymbol{\beta}}(\lambda) = \sum_{\mathbf{s}} \frac{|X_\lambda(\mathbf{s})|^2}{\sum_{\mathbf{u}} |X_\lambda(\mathbf{u})|^2} \widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s}) = \sum_{\mathbf{s}} |H_\lambda(\mathbf{s})| \widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s}) = \sum_{\mathbf{s}} \mathcal{P}_\lambda(\mathbf{s}) \widehat{\boldsymbol{\beta}}(\lambda; \mathbf{s}) \tag{4}$$

where $|\cdot|$ denotes determinant and $H_\lambda(\mathbf{s})$ is the sub-matrix of H_λ defined in (3) with row and column indices in $\mathbf{s}$. Note that $\mathcal{P}_\lambda(\mathbf{s}) = |H_\lambda(\mathbf{s})|$ is a probability measure on subsets of size p from the set $\{1, \cdots, n, n+1, \cdots, n+p\}$ and so

$$\widehat{\boldsymbol{\beta}}(\lambda) = E_\lambda[\widehat{\boldsymbol{\beta}}(\lambda; \mathbf{S})]$$

where $\mathbf{S} \sim \mathcal{P}_\lambda$. The following result is proved in Knight (2019).

Proposition 1. Suppose that $\mathbf{S} \sim \mathcal{P}_\lambda$ (as defined in (4) and $H_\lambda(\mathbf{s})$ is a sub-matrix of H_λ in (3).

(a) For $\mathbf{s} = \{i_1, \cdots, i_k\}$ where $k \leq p$,

$$P(\mathbf{s} \subset \mathbf{S}) = |H_\lambda(\mathbf{s})|.$$

(b) For an arbitrary set of indices $\mathbf{s}$, $P(\mathbf{s} \cap \mathbf{S} = \emptyset) = |I - H_\lambda(\mathbf{s})|$.

The following result was originally proved in Leamer and Chamberlain (1976). For completeness, its proof is provided in the Appendix.

Proposition 2. Define $\widehat{\boldsymbol{\beta}}_{\mathbf{j}}$ to be the OLS estimate of $\boldsymbol{\beta}$ based on predictors in $\mathbf{j} = \{j_1, \cdots, j_k\}$ (where $0 \leq k \leq p$). Then for $\widehat{\boldsymbol{\beta}}(\lambda)$ defined in (2)

$$\widehat{\boldsymbol{\beta}}(\lambda) = \sum_{\text{all } \mathbf{j}} a_\lambda(\mathbf{j}) \widehat{\boldsymbol{\beta}}_{\mathbf{j}}$$

where

$$\begin{aligned} a_\lambda(\mathbf{j}) &= \frac{|\mathcal{X}^T\mathcal{X}|}{|\mathcal{X}^T\mathcal{X} + \lambda I|} = |I - \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}| \quad \text{if } \mathbf{j} = \{1, \cdots, p\} \\ a_\lambda(\mathbf{j}) &= \frac{\lambda^p}{|\mathcal{X}^T\mathcal{X} + \lambda I|} \quad \text{if } \mathbf{j} = \emptyset \\ a_\lambda(\mathbf{j}) &= \frac{\lambda^{p-k}|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|}{|\mathcal{X}^T\mathcal{X} + \lambda I|} \quad \text{if } \text{card}(\mathbf{j}) = k \end{aligned}$$

where $\mathcal{X}(\mathbf{j})$ is the sub-matrix of $\mathcal{X}$ with column indices in $\mathbf{j}$.

2.1 Level of expression

Proposition 2 effectively says that for a given model size and fixed λ, higher weights $a_\lambda(\mathbf{j})$ are given to models with a lower degree of multicollinearity as measured by $|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|$. (Given that multicollinearity was the original *raison d'être* of ridge regression, this is not surprising.) This suggests that a given predictor will have a higher weight if it is less correlated with the other predictors. This idea can be formalized as follows: We define the **level of expression (LoE)**[1] of predictor j to be

$$\text{LoE}_\lambda(j) = \sum_{\mathbf{j}: j \in \mathbf{j}} a_\lambda(\mathbf{j}) = P(n + j \cap \mathbf{S} = \emptyset)$$

where $\mathbf{S} \sim \mathcal{P}_\lambda$. Using Proposition 1, it follows that

$$\text{LoE}_\lambda(j) = 1 - \lambda[(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}]_{jj}$$

where the subscript jj denotes the j diagonal element of the matrix. If $\text{LoE}_\lambda(j) = 1$ then the parameter estimate for predictor j is the least squares estimate of this parameter while if $\text{LoE}_\lambda(j) = 0$ then the parameter estimate is exactly 0.

In some sense, $\text{LoE}_\lambda(j)$ might be thought of as a measure of "lack of shrinkage" of the parameter estimate $\widehat{\beta}_j(\lambda)$ for predictor j relatively to its least squares estimate $\widehat{\beta}_j(0)$ (assuming that $\mathcal{X}$ has full rank); for example when $\mathcal{X}^T\mathcal{X} = I$ (or a multiple of I), we have the simple relationship

$$\widehat{\beta}_j(\lambda) = \text{LoE}_\lambda(j) \times \widehat{\beta}_j(0).$$

[1]Here we are making a crude analogy with the notion of gene expression in genetics; $\text{LoE}_\lambda(j)$ is a measure of the degree to which predictor j is "switched on" relative to least squares estimation.

More generally, LoE can be thought of as a measure of the insensitivity of $\widehat{\beta}_j(\lambda)$ to the penalty term in $\lambda\beta_j^2$. Perhaps most importantly, LoE is not, *per se*, a measure of the importance of a predictor as it relates to the response; indeed, the response may depend strongly on a predictor whose LoE for a given λ is small.

We can also extend the definition to define the LoE of a subset of predictors with indices $\mathbf{j} = \{j_1, \cdots, j_k\}$ (again using Proposition 1) as follows:

$$\text{LoE}_\lambda(\mathbf{j}) = P[(n+j_1, \cdots, n+j_k) \cap \mathbf{S} = \emptyset] = \left| I - \lambda \left[(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \right]_{\mathbf{j}} \right|$$

where $\left[(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \right]_{\mathbf{j}}$ is the $k \times k$ sub-matrix of $(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}$ with row and column indices in $\mathbf{j}$.

Some insight into the notion of LoE can be obtained by looking at approximations of $\text{LoE}_\lambda(j)$ for small and large values of λ. When λ is small and $\mathcal{X}^T\mathcal{X}$ is invertible then

$$\lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} = \lambda(\mathcal{X}^T\mathcal{X})^{-1} - \lambda^2(\mathcal{X}^T\mathcal{X}^T)^{-2} + \cdots$$

which suggests the approximation

$$\text{LoE}_\lambda(j) \approx 1 - \lambda[(\mathcal{X}^T\mathcal{X})^{-1}]_{jj}.$$

If $\mathcal{X}^T\mathcal{X}$ is a correlation matrix then the diagonal elements of $(\mathcal{X}^T\mathcal{X})^{-1}$ are variance inflation factors (VIFs) as defined in Marquardt (1970). In Example 1 below, we will consider approximating $\text{LoE}_\lambda(j)$ for small λ in the case where $\mathcal{X}^T\mathcal{X}$ is not invertible.

Conversely, for large λ, an approximation for $\text{LoE}_\lambda(j)$ can be obtained from the expansion

$$\lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \approx I - \frac{1}{\lambda}\mathcal{X}^T\mathcal{X} + \frac{1}{\lambda}(\mathcal{X}^T\mathcal{X})^2.$$

Thus if $\mathcal{X}^T\mathcal{X}$ is a correlation matrix

$$\begin{aligned} \text{LoE}_\lambda(j) &\approx \frac{1}{\lambda} - \frac{1}{\lambda^2}[(\mathcal{X}^T\mathcal{X})^2]_{jj} \\ &= \frac{1}{\lambda} - \frac{1}{\lambda^2}\sum_{k=1}^{p}\rho_{jk}^2 \end{aligned}$$

where $\rho_{jk} = [\mathcal{X}^T\mathcal{X}]_{jk}$ is the correlation between predictors j and k. This latter approximation does not require $\mathcal{X}^T\mathcal{X}$ to be invertible.

Proposition 2 holds even if the rank of $\mathcal{X}$ is less than p, for example, if $p > n$. If $\text{rank}(\mathcal{X}) = r < p$ then for any $\mathbf{j}$ with $\text{card}(\mathbf{j}) \geq r+1$, we have $a_\lambda(\mathbf{j}) = 0$. Thus $\widehat{\boldsymbol{\beta}}(\lambda)$ is determined by OLS estimates based on subsets of r or fewer predictors. Moreover, unlike the full rank case where the LoE of each predictor tends tends to 1 as $\lambda \to 0$, we obtain a much more interesting limiting behaviour for $\text{LoE}_\lambda(j)$ in the ridgeless case (Hastie *et al.*, 2019) for $r < p$ as shown in the following example.

Example 1. If $\text{rank}(\mathcal{X}) = r < p$ then as $\lambda \to 0$, $a_\lambda(\mathbf{j}) \to 0$ when $\text{card}(\mathbf{j}) < r$ while if $\text{card}(\mathbf{j}) = r$, we have

$$a_\lambda(\mathbf{j}) = \frac{\lambda^{p-r}|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|}{|\mathcal{X}^T\mathcal{X} + \lambda I|};$$

the denominator can be written as

$$|\mathcal{X}^T\mathcal{X} + \lambda I| = \lambda^{p-r}\prod_{j=1}^{r}(\lambda + \mu_j)$$

$\mu_1 \geq \cdots \geq \mu_r$ are the positive eigenvalues of $\mathcal{X}^T\mathcal{X}$. From this, as $\lambda \to 0$

$$a_\lambda(\mathbf{j}) \to \frac{|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|}{\mu_1 \times \cdots \times \mu_r} = a_0(\mathbf{j})$$

for subsets $\mathbf{j}$ of size r. Denoting $\widehat{\boldsymbol{\beta}}(0)$ as the ridgeless estimate, we have

$$\widehat{\boldsymbol{\beta}}(0) = \sum_{\text{card}(\mathbf{j})=r} a_0(\mathbf{j})\widehat{\boldsymbol{\beta}}_\mathbf{j}).$$

As before, we can define the LoE of predictor j in the ridgeless estimate:

$$\text{LoE}_0(j) = \sum_{\mathbf{j}:j\in\mathbf{j}} a_0(\mathbf{j}).$$

$\text{LoE}_0(j)$ can be determined from the decomposition $\mathcal{X}^T\mathcal{X} = \Gamma_+ D_+ \Gamma_+^T$ where Γ_+ is a $p \times r$ matrix whose columns are the orthonormal eigenvectors corresponding to positive eigenvalues (contained in the $r \times r$ diagonal matrix D_+): As $\lambda \to 0$,

$$I - \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \to \Gamma_+\Gamma_+^T$$

since the eigenvalues of $I - \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}$ tend to 1 or 0 depending on whether the corresponding eigenvalue of $\mathcal{X}^T\mathcal{X}$ is positive or 0. Thus if $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p$ are the rows of Γ_+ then

$$\begin{aligned}\text{LoE}_0(j) &= \lim_{\lambda\to 0} 1 - \lambda[(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}]_{jj} \\ &= \|\boldsymbol{u}_j\|^2\end{aligned}$$

In terms of principal components (PCs), $\text{LoE}_0(j)$ is simply the sum of squared correlations between predictor j and the r PCs with positive eigenvalues. (Geometrically, if ϕ is the angle between the j coordinate vector and the null space of $\mathcal{X}$ then $\text{LoE}_0(j) = \sin^2(\phi)$.) The matrix $\Gamma_+\Gamma_+^T$ can also be used to evaluate $\text{LoE}_0(\mathbf{j})$ for a collection of predictors. For example, for $\mathbf{j} = \{j_1, j_2\}$, we have

$$\text{LoE}_0(\mathbf{j}) = \|\boldsymbol{u}_{j_1}\|^2\|\boldsymbol{u}_{j_2}\|^2 - (\boldsymbol{u}_{j_1}^T\boldsymbol{u}_{j_2})^2.$$

In the case where $\mathcal{X}^T\mathcal{X}$ is invertible, we noted the relationship between LoE and VIF when λ is close to 0. In the ridgeless case, we also have a small λ approximation:

$$\text{LoE}_\lambda(j) \approx \|\boldsymbol{u}_j\|^2 - \lambda\boldsymbol{u}_j D_+^{-1}\boldsymbol{u}_j^T = \|\boldsymbol{u}_j\|^2 - \lambda\sum_{h=1}^{r}\mu_h^{-1}u_{jh}^2$$

This representation suggests a two-dimensional analogue of the VIF in the case where $\mathcal{X}$ is non-full rank. If $\mathcal{X}$ is normalized so that $\mathcal{X}^T\mathcal{X}$ is a correlation matrix then we can define the VIF for predictor j as follows:

$$\mathrm{VIF}_0(j) = \left(\|\boldsymbol{u}_j\|^2, \sum_{h=1}^{r} \mu_h^{-1} u_{jh}^2 \right).$$

The first term of $\mathrm{VIF}_0(j)$ is simply $\mathrm{LoE}_0(j)$ while the second term is a variance ratio comparing the variance of the ridgeless estimator of β_j for $\mathcal{X}^T\mathcal{X} = \Gamma_+ D_+ \Gamma_+^T$ to the variance of an estimator of β_j assuming orthogonality of predictors. In the full rank case,

$$\mathrm{VIF}_0(j) = (1, \mathrm{VIF}(j))$$

where $\mathrm{VIF}(j)$ is the standard definition of VIF for predictor j. In the rank deficient case ($r < p$), the second component of $\mathrm{VIF}_0(j)$ can be less than 1, which is a consequence of shrinkage.

How do we interpret the values of $\{\mathrm{LoE}_0(j)\}$ and $\{\mathrm{VIF}_0(j)\}$? As mentioned above, $\mathrm{LoE}_0(j) = \sin^2(\phi)$ where ϕ is the angle between the coordinate vector $\boldsymbol{e}_j$ and the null space of $\mathcal{X}^T\mathcal{X}$. So roughly speaking, the more that predictor j is correlated with the other $p-1$ predictors, the smaller $\mathrm{LoE}_0(j)$ will be. Moreover, from the definition of $\{\mathrm{LoE}_0(j)\}$, we have

$$\sum_{j=1}^{p} \mathrm{LoE}_0(j) = r$$

and so if r/p is small, we will (inevitably) have $\mathrm{LoE}_0(j)$ close to 0 for some j. We also have the lower bound

$$\mathrm{LoE}_0(j) \geq \frac{[\mathcal{X}^T\mathcal{X}]_{jj}}{\mu_1};$$

this follows from the Cauchy-Schwarz inequality, noting that if $[\mathcal{X}^T\mathcal{X}]_{jj} = \boldsymbol{v}^T D_+ \boldsymbol{v}$ then

$$\begin{aligned} [\mathcal{X}^T\mathcal{X}]_{jj} &= \boldsymbol{v}^T D_+ \boldsymbol{v} \\ &\leq (\boldsymbol{v}^T D_+^2 \boldsymbol{v})^{1/2} \|\boldsymbol{v}\| \\ &\leq \mu_1 \|\boldsymbol{v}\|^2. \end{aligned}$$

The lower bound can be attained: If $\mathcal{X}^T\mathcal{X}$ is a correlation matrix and $\boldsymbol{u}_j = (1/\sqrt{\mu_1}, 0, \cdots, 0)$ then $\mathrm{LoE}_0(j) = 1/\mu_1$. This scenario implies that predictor j is weakly correlated with the first PC and uncorrelated with the remaining $r-1$ PCs with positive eigenvalues. Likewise, we have the following bound for the two components of $\mathrm{VIF}_0(j)$:

$$\left\{ \sum_{h=1}^{r} \mu_h^{-1} u_{jh}^2 \right\}^{1/2} \geq \|\boldsymbol{u}_j\|^2,$$

where the bound is attained (as before) if $\boldsymbol{u}_j = (1/\sqrt{\mu_1}, 0, \cdots, 0)$.

In practice, we would like to identify predictors whose LoE is large or small relative to some reference distribution for $\{\mathrm{LoE}_0(j)\}$; for example, we may want to retain only those predictors with $\mathrm{LoE}_0(j)$ greater than some threshold. Suppose that $\mathcal{X}^T\mathcal{X}$ is a correlation matrix so that $\mathrm{LoE}_0(j) \geq \mu_1^{-1}$ for all j and suppose that Γ is a uniformly distributed (random) orthogonal matrix; the marginal distribution of each row (or of each column) of Γ can be represented by the random vector $\boldsymbol{Z}/\|\boldsymbol{Z}\|_2$ where $\boldsymbol{Z} = (Z_1, \cdots, Z_p) \sim \mathcal{N}_p(\boldsymbol{0}, I)$ and the random vector

$$\left(\frac{Z_1^2}{\|\boldsymbol{Z}\|^2}, \cdots, \frac{Z_p^2}{\|\boldsymbol{Z}\|^2}\right)$$

has a Dirichlet distribution with concentration parameter $(1/2, \cdots, 1/2)$. We know that $\mathrm{LoE}_0(j) \geq \mu_1^{-1}$ so that the distribution of $(Z_1^2 + \cdots + Z_r^2)/\|\boldsymbol{Z}\|^2$ is concentrated on the interval $[\mu_1^{-1}, 1]$. If we condition on $Z_1^2/\|\boldsymbol{Z}\|^2 = \mu_1^{-1}$ then the conditional distribution of $(Z_1^2 + \cdots + Z_r^2)/\|\boldsymbol{Z}\|^2$ is concentrated on the interval $[\mu_1^{-1}, 1]$ and this distribution can be represented as $(1 - \mu_1^{-1})W + \mu_1^{-1}$ where W has a $\mathrm{Beta}((r-1)/2, (p-r)/2)$ distribution.

If we pretend that $\{\mathrm{LoE}_0(j)\}$ are independent random variables with this distribution (thereby ignoring the correlation between them) then we might think of using extreme quantiles of the Beta distribution (appropriately normalized) as upper and lower thresholds for $\|\mathrm{LoE}_0(j)\}$; for example, if q_τ and $q_{1-\tau}$ are quantiles of the $\mathrm{Beta}((r-1)/2, (p-r)/2)$ distribution (where τ is close to 0) then we can use $(1-\mu_1^{-1})q_\tau + \mu_1^{-1}$ and $(1-\mu_1^{-1})q_{1-\tau} + \mu_1^{-1}$ as lower and upper thresholds.. If r and p are reasonably large and $\alpha = (r-1)/(p-1)$ is bounded away from 0 and 1 then the distribution of $(1 - \mu_1^{-1})W + \mu_1^{-1}$ can be approximated by a Normal distribution with mean $(1 - \mu_1^{-1})\alpha + \mu_1^{-1}$ and variance $2(1 - \mu_1^{-1})^2\alpha(1-\alpha)/(p+1)$, which yields the upper and lower thresholds:

$$(1 - \mu_1^{-1})\alpha + \mu_1^{-1} \pm 2(1 - \mu_1^{-1})\sqrt{\alpha(1-\alpha)\ln(p)/(p+1)}.$$

These thresholds allow us to identify predictors whose LoEs are either larger or smaller than what one might expect given the simple uniform model on orthogonal matrices.

The relationship between the two components of $\mathrm{VIF}_0(j)$ for $j = 1, \cdots, p$ is more complex and will not be explored in depth here. Typically, there will be a positive correlation between the two components, with the correlation increasing as r/p decreases. A simple (somewhat naive) "reference" correlation can be obtained from the 'Dirichlet model described above:

$$\mathrm{Corr}\left(\|\boldsymbol{u}_j\|^2, \sum_{h=1}^{r} \mu_h^{-1} u_{jh}^2\right) \approx \left(1 - \frac{r}{p}\right)^{1/2} \left\{\frac{1}{r}\sum_{h=1}^{r}\left(\frac{\mu_h^{-2}}{\nu^2} - \frac{r}{p}\right)\right\}^{-1/2}$$

where $\nu = (\mu_1^{-1} + \cdots + \mu_r^{-1})/r$. As $r/p \to 1$, this correlation tends to 0 while the limit as $r/p \to 0$ is $[(\mu_1^{-2}/\nu^2 + \cdots + \mu_r^{-2}/\nu^2)/r]^{-1/2}$, which is less than 1 unless $\mu_1 = \cdots = \mu_r = p/r$.

When is $\mathrm{LoE}_0(j) = 1$? Suppose that $\mathcal{X}$ has the form

$$\mathcal{X} = (\mathcal{X}_1 \ \mathcal{X}_2 B)$$

so that

$$\mathcal{X}^T\mathcal{X} = \begin{pmatrix} \mathcal{X}_1^T\mathcal{X}_1 & \mathcal{X}_1^T\mathcal{X}_2 B \\ B^T\mathcal{X}_2^T\mathcal{X}_1 & B^T\mathcal{X}_2^T\mathcal{X}_2 B \end{pmatrix}$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ are, respectively, $n \times r_1$ and $n \times r_2$ matrices whose $r = r_1 + r_2$ columns are linearly independent and B is a $r_2 \times (p - r_1)$ matrix whose rank is at least r_2. In this case, $\mathrm{LoE}_0(j) = 1$ for $j = 1, \cdots, r_1$. This follows since $|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|$ (and hence $a_0(\mathbf{j})$) is positive only if $\{1, \cdots, r_1\} \subset \mathbf{j}$. This also follows from our definition: If $\boldsymbol{v}$ is an eigenvector of $\mathcal{X}^T\mathcal{X}$ with eigenvalue 0 then the first r_1 elements of $\boldsymbol{v}$ are 0; therefore, the squared norms of the rows $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{r_1}$ of Γ_+ must equal 1.

The result of Proposition 2 also holds if we replace the ridge penalty $\lambda \sum_{j=1}^p \beta_j^2$ by $\sum_{j=1}^p \lambda_j \beta_j^2 = \boldsymbol{\beta}^T \Lambda \boldsymbol{\beta}$ for $\lambda_1, \cdots, \lambda_p \geq 0$ where we replace $|\mathcal{X}^T\mathcal{X} + \lambda I|$ by $|\mathcal{X}^T\mathcal{X} + \Lambda|$ and λ^{p-k} by $\prod_{h \notin \mathbf{j}} \lambda_h$. Likewise, if the penalty is $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}$ for some symmetric positive definite matrix $\Sigma = L^T L$, we can replace $\mathcal{X}$ in Proposition 2 by $\mathcal{X}L^{-1}$ with $\lambda = 1$. A judicious choice of L will give the j column of $\mathcal{X}L^{-1}$ equal to a multiple of the j column of $\mathcal{X}$; this is useful for defining LoE in this general case.

Defining $\mathrm{LoE}(j)$ when the penalty term has the general form $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}$ requires some thought particularly when $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}$ is not additively separable. If Σ is diagonal with diagonal elements $\sigma_{11}, \cdots, \sigma_{pp}$ then the extension is straightforward:

$$\mathrm{LoE}_\Sigma(j) = 1 - \sigma_{jj}[(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}]_{jj}$$

which follows directly from Proposition 1.

In the case where Σ is positive definite, we have $\Sigma = L^T L$ where L is non-unique; to define $\mathrm{LoE}_\Sigma(j)$, we need to find $L = L_j$ (essentially a reparametrization) such that the j component of $\boldsymbol{\theta} = L_j \boldsymbol{\beta}$ is a multiple of β_j:

$$\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} = \sum_{h \neq j} (\boldsymbol{\ell}_h^T \boldsymbol{\beta})^2 + (\boldsymbol{\ell}_j^T \boldsymbol{\beta})^2 = \sum_{h \neq j} \theta_h^2 + \ell_{jj}^2 \beta_j^2$$

where $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_p$ are linearly independent vectors with $\boldsymbol{\ell}_j = \ell_{jj} \boldsymbol{e}_j$, a multiple of the j coordinate vector $\boldsymbol{e}_j$ so that $(\boldsymbol{\ell}_j^T \boldsymbol{\beta})^2 = \ell_{jj}^2 \beta_j^2$. The existence of L_j follows from a Cholesky factorization of Σ with a reordering of rows and columns so that $\Sigma = L_j^T L_j$ where the j row of L_j is the j coordinate vector $\boldsymbol{e}_j$ multiplied by ℓ_{jj}. From this, it follows that

$$\mathrm{LoE}_\Sigma(j) = 1 - \ell_{jj}^2 [(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}]_{jj}$$

where we apply Proposition 1 using the projection matrix

$$H_j = \begin{pmatrix} \mathcal{X}(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}\mathcal{X}^T & \mathcal{X}(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}L_j^T \\ L_j(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}\mathcal{X}^T & L_j(\mathcal{X}^T\mathcal{X} + \Sigma)^{-1}L_j^T \end{pmatrix}.$$

Applications of this are given in Examples 2 and 3 below.

In order to compute $\text{LoE}_\Sigma(j)$, we do not need to compute the matrix L_j, only ℓ_{jj}^2. A simple approach to determining ℓ_{jj}^2 uses the fact that if $\Sigma = L^T L$ then for any orthogonal $p \times p$ matrix O, we have $\Sigma = (OL)^T(OL)$. Thus for a given L, we simply need to find $O = O_j$ so that the j row of $O_j L$ is a multiple of $\boldsymbol{e}_j$; in fact, we need only determine the j row of O_j, which can be determined by solving the equation

$$L^T \boldsymbol{a}_j = \boldsymbol{e}_j$$

with $\ell_{jj}^2 = 1/\|\boldsymbol{a}_j\|^2$. (The j row of O_j is $\boldsymbol{a}_j^T/\|\boldsymbol{a}_j\|$ with the remaining rows of O_j determined via orthogonalization.) Note that ℓ_{jj}^2 does not depend on the choice of L; if we replace L by OL for some orthogonal matrix O then ℓ_{jj}^2 remains unchanged.

(Using similar arguments, we can define $\text{LoE}_\Sigma(\mathbf{j})$ for a subset of predictors $\mathbf{j}$ by

$$\text{LoE}_\Sigma(\mathbf{j}) = \left| I - \left[L_{\mathbf{j}} (\mathcal{X}^T \mathcal{X} + \Sigma)^{-1} L_{\mathbf{j}}^T \right]_{\mathbf{j}} \right|$$

where again we use a Cholesky factorization of Σ, reordering rows and columns to obtain $L_{\mathbf{j}}$ such that $L_{\mathbf{j}}^T L_{\mathbf{j}} = \Sigma$.)

If Σ is singular and non-negative definite (with $\mathcal{X}^T\mathcal{X} + \Sigma$ non-singular) then the approach used when Σ is positive definite is not necessarily applicable. However, the basic idea for defining $\text{LoE}_\Sigma(j)$ described above still applies with an additional wrinkle.

Suppose that the rank of Σ is $r < p$; then $\Sigma = L^T L$ where L is $r \times p$ and L is determined up to $r \times r$ orthogonal transformation O ($(OL)^T(OL) = L^T L$). If there exists an orthogonal matrix O_j such that the j row of $O_j L$ is a multiple of the coordinate vector $\boldsymbol{e}_j$ (that is, $\boldsymbol{e}_j$ lies in the column space of L^T) then as before

$$\text{LoE}_\Sigma(j) = 1 - \ell_{jj}^2 [(\mathcal{X}^T \mathcal{X} + \Sigma)^{-1}]_{jj}$$

where $\ell_{jj}^2 = 1/\|\boldsymbol{a}_j\|^2$ with $L^T \boldsymbol{a}_j = \boldsymbol{e}_j$. If no such orthogonal matrix exists (that is, $L^T \boldsymbol{a}_j = \boldsymbol{e}_j$ does not have a solution) then $\text{LoE}_\Sigma(j) = 1$; in this case, if

$$\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} = \sum_{h \neq j} (\boldsymbol{\ell}_h^T \boldsymbol{\beta})^2 + \ell_{jj}^2 \beta_j^2$$

then $\ell_{jj}^2 = 0$. For example, if $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} = (\mathbf{1}^T \boldsymbol{\beta})^2$ (where $\mathbf{1}$ is a vector of 1s) then $\text{LoE}_\Sigma(j) = 1$ for all j; this follows from Proposition 1 as the constraint $\mathbf{1}^T \boldsymbol{\beta} = 0$ does not guarantee any element of $\boldsymbol{\beta}$ to be exactly 0. However, if we reparametrize so that $\boldsymbol{\theta} = A\boldsymbol{\beta}$ where A is non-singular and $\theta_1 = \mathbf{1}^T \boldsymbol{\beta}$ then LoE for the new first predictor (the first column of $\mathcal{X} A^{-1}$ will be less than 1 (with the remaining LoEs equal to 1).

The following example applies the theory developed above to boosted ridge regression.

Example 2. Consider "boosting" ridge regression by the following iterative proces: If $\widehat{\boldsymbol{\beta}}^{(k-1)}(\lambda)$ is our estimate of $\boldsymbol{\beta}$ at step $k-1$ of boosting then the estimate at step k is given

by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(k)}(\lambda) &= \widehat{\boldsymbol{\beta}}^{(k-1)}(\lambda) + (\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T\left\{\boldsymbol{y} - \mathcal{X}\widehat{\boldsymbol{\beta}}^{(k-1)}(\lambda)\right\} \\ &= \sum_{h=1}^{k}\lambda^{h-1}\left(\mathcal{X}^T\mathcal{X} + \lambda I\right)^{-h}\mathcal{X}^T\mathbf{y}\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}^{(0)}(\lambda) = \mathbf{0}$. It follows that $\widehat{\boldsymbol{\beta}}^{(k)}(\lambda)$ minimizes

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \boldsymbol{\beta}^T\Sigma_k\boldsymbol{\beta}$$

where

$$\Sigma_k = \lambda\left\{\sum_{h=1}^{k}\lambda^h\left(\mathcal{X}^T\mathcal{X} + \lambda I\right)^{-h}\right\}^{-1} - \mathcal{X}^T\mathcal{X}.$$

For $k \geq 2$, Σ_k is a diagonal matrix if, and only if, $\mathcal{X}^T\mathcal{X}$ is a diagonal matrix.

To evaluate the LoE for a given predictor, define $r \leq p$ to be the rank of $\mathcal{X}$ and write $\mathcal{X}^T\mathcal{X} = \Gamma D\Gamma^T = \Gamma_+ D_+\Gamma_+^T$ where D and D_+ are diagonal matrices whose elements are, respectively, the eigenvalues and positive eigenvalues $(\mu_1, \cdots, \mu_r > 0)$ of $\mathcal{X}^T\mathcal{X}$ and $\Gamma = (\Gamma_+\ \Gamma_0)$ whose columns are orthonormal eigenvectors of $\mathcal{X}^T\mathcal{X}$. If $(u_{j1}, \cdots, u_{jp})$ is the j row of Γ then

$$\mathrm{LoE}_{\Sigma_k}(j) = 1 - \ell_{jj}^2(\lambda, k)\left\{\sum_{h=1}^{r}\mu_h^{-1}\left[1 - \left(\frac{\lambda}{\lambda + \mu_h}\right)^k\right]u_{jh}^2 + \frac{k}{\lambda}\sum_{h=r+1}^{p}u_{jh}^2\right\}$$

where

$$\ell_{jj}^2(\lambda, k) = \left\{\sum_{h=1}^{r}\mu_h^{-1}\left[(1 + \mu_h/\lambda)^k - 1\right]u_{jh}^2 + \frac{k}{\lambda}\sum_{h=r+1}^{p}u_{jh}^2\right\}^{-1}.$$

When both k and λ are large so that $k/\lambda = \tau$, we can approximate $\mathrm{LoE}_{\Sigma_k}(j)$ as follows:

$$\mathrm{LoE}_{\Sigma_k}(j) \approx 1 - \frac{\sum_{h=1}^{r}\mu_h^{-1}\left[1 - \exp(-\tau\mu_h)\right]u_{jh}^2 + \tau\sum_{h=r+1}^{p}u_{jh}^2}{\sum_{h=1}^{r}\mu_h^{-1}\left[\exp(\tau\mu_h) - 1\right]u_{jh}^2 + \tau\sum_{h=r+1}^{p}u_{jh}^2}.$$

As $\tau \to \infty$, $\mathrm{LoE}_{\Sigma_k}(j) \to 1$ unless $\sum_{h=r+1}^{p}u_{jh}^2 = 1$, in which case $\mathrm{LoE}_{\Sigma_k}(j) = 0$.

The idea of ridgeless regression (for $r < p$) in Example 1 can be extended to "ridgeless boosting" whereby we take $\lambda \to 0$ for some fixed $k \geq 2$. For λ close to 0, we have

$$\ell_{jj}^2(\lambda, k) = \frac{\lambda}{k} + o(\lambda).$$

Thus as $\lambda \to 0$,

$$\mathrm{LoE}_{\Sigma_k}(j) \to 1 - \left\{\sum_{h=r+1}^{p}u_{jh}^2\right\}^2 = \sum_{h=1}^{r}u_{jh}^2,$$

which is the same as for ridgeless regression ($k = 1$) as shown in Example 1. This suggests that ridgeless regression and ridgeless boosting produce the same estimates; this is (perhaps) not surprising and is straightforward to show mathematically. For a given λ, $\widehat{\boldsymbol{\beta}}^{(k)}(\lambda)$ minimizes

$$Z_\lambda^{(k)}(\boldsymbol{\beta}) = \frac{1}{\lambda^k}\left\{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \eta^2\right\} + \frac{1}{\lambda^k}\boldsymbol{\beta}^T\Sigma_k\boldsymbol{\beta}$$

where η^2 is the minimum value of $\sum(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$; we also have

$$\Sigma_k = \frac{\lambda}{k}\Gamma_0\Gamma_0^T + \lambda^k\Gamma_+ D_+^{-(k-1)}\Gamma_+^T + o(\lambda^k).$$

Now $Z_\lambda^{(1)}$ converges pointwise to $Z^{(1)}$ where

$$Z^{(1)}(\boldsymbol{\beta}) = \begin{cases} \boldsymbol{\beta}^T\boldsymbol{\beta} & \text{if } \mathcal{X}^T\mathcal{X}\boldsymbol{\beta} = \mathcal{X}^T\boldsymbol{y} \\ +\infty & \text{otherwise} \end{cases}$$

while for $k \geq 2$, $Z_\lambda^{(k)}$ converges pointwise to $Z^{(k)}$ where

$$Z^{(k)}(\boldsymbol{\beta}) = \begin{cases} \boldsymbol{\beta}^T\Gamma_+ D_+^{-(k-1)}\Gamma_+^T\boldsymbol{\beta} & \text{if } \mathcal{X}^T\mathcal{X}\boldsymbol{\beta} = \mathcal{X}^T\boldsymbol{y} \text{ and } \Gamma_0^T\boldsymbol{\beta} = \mathbf{0} \\ +\infty & \text{otherwise.} \end{cases}$$

The pointwise convergence can be turned into epi-convergence (Attouch and Wets, 1981), which (together with convexity) guarantees convergence of the minimizers as $\lambda \to 0$. Although the limiting functions $\{Z^{(k)}(\boldsymbol{\beta})\}$ are different, their minimizers are the same:

$$\widehat{\boldsymbol{\beta}}^{(k)}(\lambda) \to \widehat{\boldsymbol{\beta}}(0) = \Gamma_+ D_+^{-1}\Gamma_+^T\mathcal{X}^T\boldsymbol{y}.$$

For $k \geq 2$, this follows from the constraint $\Gamma_0^T\boldsymbol{\beta} = 0$ while for $k = 1$, it follows from the fact that the minimizer of $\boldsymbol{\beta}^T\boldsymbol{\beta}$ over all OLS estimates must satisfy $\boldsymbol{\beta}^T\boldsymbol{a} = 0$ for all $\boldsymbol{a}$ in the null space of $\mathcal{X}^T\mathcal{X}$; hence the constraint $\Gamma_0^T\boldsymbol{\beta} = \mathbf{0}$ holds automatically. What distinguishes $\{\widehat{\boldsymbol{\beta}}^{(k)}(\lambda)\}$ for different k is the path the estimates take to $\widehat{\boldsymbol{\beta}}(0)$: As $\lambda \to 0$, we have

$$\lambda^{-k}(\widehat{\boldsymbol{\beta}}^{(k)}(\lambda) - \widehat{\boldsymbol{\beta}}(0)) \to -\Gamma_+ D_+^{-(k+1)}\Gamma_+^T\mathcal{X}^T\boldsymbol{y}.$$

While this would suggest a faster convergence rate (as $\lambda \to 0$) for larger k, this faster convergence would also be mitigated somewhat if $\mathcal{X}^T\mathcal{X}$ has small but positive eigenvalues.

The following example shows how we might apply the notion of LoE for principal component (PC) and approximate PC regression.

Example 3. PC regression is a commonly used method for dimension reduction. The idea is to use the eigen-decomposition of $\mathcal{X}^T\mathcal{X} = \Gamma D\Gamma^T$ to define a collection of d uncorrelated predictors $\mathcal{X}_* = \mathcal{X}\Gamma_*$ where Γ_* is a $p \times d$ sub-matrix of Γ; each of these d predictors is a function of the original set of predictors. For convenience, we will assume that Γ_* contains

the first d columns of Γ and the eigenvalues $\mu_1, \cdots, \mu_p$ of $\mathcal{X}^T\mathcal{X}$ satisfy $\mu_1, \cdots, \mu_d > 0$. For PC regression $\mu_1, \cdots, \mu_d$ are the d largest eigenvalues although the following discussion will not depend on this.

The PC estimate is given by

$$\widehat{\boldsymbol{\theta}}_* = (\mathcal{X}_*^T\mathcal{X}_*)^{-1}\mathcal{X}_*^T\boldsymbol{y}$$

and the "implied" estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_* = \Gamma_*\widehat{\boldsymbol{\theta}}_*$. We can define $\widehat{\boldsymbol{\beta}}_*$ as the limit of a sequence of generalized ridge regression estimates. Specifically, define $\widehat{\boldsymbol{\beta}}_\eta$ to minimize

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \boldsymbol{\beta}^T\Sigma_\eta\boldsymbol{\beta}$$

where

$$\Sigma_\eta = \Gamma V_\eta \Gamma^T$$

with V_η a diagonal matrix whose first d elements are $\eta_1, \cdots, \eta_d$ and last $p-d$ elements are $1/\eta_{d+1}, \cdots, 1/\eta_p$. As $\eta_1, \cdots, \eta_p \to 0$, $\widehat{\boldsymbol{\beta}}_\eta \to \widehat{\boldsymbol{\beta}}_*$. This particular formulation of the PC estimate $\widehat{\boldsymbol{\beta}}_*$ is more than a simple mathematical device. For example, if p is very large, computing an eigen-decomposition of $\mathcal{X}^T\mathcal{X}$ may be computationally expensive and so approximating $\widehat{\boldsymbol{\beta}}_*$ by some sort of $\widehat{\boldsymbol{\beta}}_\eta$ may be useful; an example of this approach is given in section 6.2.

As in Example 2, the structure of the objective function greatly facilitates the computation of the LoE for each $(\eta_1, \cdots, \eta_p)$ since $\mathcal{X}^T\mathcal{X}$ and Σ_η have the same eigen-structure:

$$\mathcal{X}^T\mathcal{X} + \Sigma_\eta = \Gamma(D + V_\eta)\Gamma^T$$

so that

$$[(\mathcal{X}^T\mathcal{X} + \Sigma_\eta)^{-1}]_{jj} = \sum_{h=1}^{d}(\mu_h + \eta_h)^{-1}u_{jh}^2 + \sum_{h=d+1}^{p}(\mu_h + \eta_h^{-1})^{-1}u_{jh}^2$$

where $(u_{j1}, \cdots, u_{jp})$ is the j row of Γ. Likewise, we can define $\ell_{jj}^2 = 1/\|\boldsymbol{a}_j\|^2$ where $\Gamma V_\eta^{1/2}\boldsymbol{a}_j = \boldsymbol{e}_j$ so that $\boldsymbol{a}_j = V_\eta^{-1/2}\Gamma^T\boldsymbol{e}_j$ and

$$\ell_{jj}^2 = \left\{\sum_{h=1}^{d}\eta_h^{-1}u_{jh}^2 + \sum_{h=d+1}^{p}\eta_h u_{jh}^2\right\}^{-1}.$$

Thus for $\eta_1, \cdots, \eta_p > 0$, we have

$$\begin{aligned}\mathrm{LoE}_\eta(j) &= 1 - \frac{\sum_{h=1}^{d}(\mu_h + \eta_h)^{-1}u_{jh}^2 + \sum_{h=d+1}^{p}(\mu_h + \eta_h^{-1})^{-1}u_{jh}^2}{\sum_{h=1}^{d}\eta_h^{-1}u_{jh}^2 + \sum_{h=d+1}^{p}\eta_h u_{jh}^2} \\ &= 1 - \sum_{h=1}^{d}\eta_h(\mu_h + \eta_h)^{-1}\mathcal{Q}_\eta(h) - \sum_{h=d+1}^{p}\eta_h^{-1}(\mu_h + \eta_h^{-1})^{-1}\mathcal{Q}_\eta(h)\end{aligned}$$

where $\mathcal{Q}_\eta(h)$ is a probability measure on $\{1, \cdots, p\}$ whose probability mass (as $\eta_1, \cdots, \eta_p \to 0$) becomes concentrated on $\{1, \cdots, d\}$ if $\sum_{h=1}^d u_{jh}^2 > 0$ and concentrated on $\{d+1, \cdots, p\}$ if $\sum_{h=1}^d u_{jh}^2 = 0$. Defining $\mathrm{LoE}_*(j)$ is the limit of $\mathrm{LoE}_\eta(j)$ as $\eta_1, \cdots, \eta_p \to 0$, we have

$$\mathrm{LoE}_*(j) = \begin{cases} 1 & \text{if } \sum_{h=1}^d u_{jh}^2 > 0 \\ 0 & \text{if } \sum_{h=1}^d u_{jh}^2 = 0. \end{cases}$$

This result is somewhat underwhelming but, upon reflection, not at all surprising — the PC estimate is a least squares estimate and will depend on predictor j provided that at least one of the PCs depends on predictor j. Some additional insight main be gained by looking at $\mathrm{LoE}_\eta(j)$ for values of $\eta_1, \cdots, \eta_p$ close to 0. For example, if $\eta_1 = \cdots = \eta_p = \eta_0$ then

$$\mathrm{LoE}_\eta(j) \approx 1 - \eta_0 \left\{ \sum_{h=1}^d u_{jh}^2 \right\}^{-1} \left\{ \sum_{h=1}^d \mu_h^{-1} u_{jh}^2 \right\} \quad \text{when } \sum_{h=1}^d u_{jh}^2 > 0$$

and

$$\mathrm{LoE}_\eta(j) \approx \eta_0 \sum_{h=d+1}^p \mu_h \quad \text{when } \sum_{h=1}^d u_{jh}^2 = 0.$$

The former case is the more interesting; For a given predictor j, $\mathrm{LoE}_\eta(j)$ is closest to 1 for small η_0 if this predictor is correlated with the PC having the largest eigenvalue and uncorrelated with the remaining $d-1$ PCs in the model, even if this correlation is very weak. In essence, by virtue of being correlated (however slightly) with the first PC, this predictor is able to piggyback on the importance of the first PC in approximate PC regression.

3 Decomposing degrees of freedom

The notion of degrees of freedom or effective degrees of freedom as a measure of model complexity follows from the work of Stein (1981) and Efron (1986). Given a response vector (of length n) $\boldsymbol{Y}$ with mean vector is $\boldsymbol{\mu}$ and covariance matrix $\sigma^2 I$, suppose that $\widehat{\boldsymbol{Y}}$ is an estimator of $\boldsymbol{\mu}$, the so-called fitted values. Efron shows that the degrees of freedom associated with $\widehat{\boldsymbol{Y}}$ is

$$\mathrm{df}(\widehat{\boldsymbol{Y}}) = \frac{1}{\sigma^2} \sum_{i=1}^n \mathrm{Cov}(\widehat{Y}_i, Y_i).$$

In the case where $\widehat{\boldsymbol{Y}} = A\boldsymbol{Y}$, we have $\mathrm{df}(\widehat{\boldsymbol{Y}}) = \mathrm{tr}(A)$; for ridge regression,

$$\mathrm{df}(\widehat{\boldsymbol{Y}}) = \sum_{j=1}^p \frac{\mu_j}{\mu_j + \lambda}$$

where $\mu_1, \cdots, \mu_p$ are the eigenvalues of $\mathcal{X}^T\mathcal{X}$. However, using $\mathrm{df}(\widehat{\boldsymbol{Y}})$ as a measure of model size is somewhat problematic; for example, Janson *et al.* (2015) give a critique of the notion of effective degrees of freedom when $\widehat{\boldsymbol{Y}}$ is a non-linear function of $\boldsymbol{Y}$.

In ridge regression, Proposition 2 illustrates that if $\mathrm{df}(\widehat{\boldsymbol{Y}}) = r < p$ then $\widehat{\boldsymbol{Y}}$ will be a convex combination of 0 through r parameter models. For example, if $\mathcal{X}^T\mathcal{X} = I$ and $\mathbf{j} = \{j_1, \cdots, j_k\}$ then

$$a_\lambda(\mathbf{j}) = \left(\frac{1}{1+\lambda}\right)^k \left(\frac{\lambda}{1+\lambda}\right)^{p-k}$$

and so

$$\sum_{\mathrm{card}(\mathbf{j})=k} a_\lambda(\mathbf{j}) = \binom{p}{k} \left(\frac{1}{1+\lambda}\right)^k \left(\frac{\lambda}{1+\lambda}\right)^{p-k}$$

for $k = 0, \cdots, p$. The following result (which is a refinement of Proposition 2) generalizes this special case and shows that $\mathrm{df}(\widehat{\boldsymbol{Y}})$ can be expressed as the expected value of a random variable N whose distribution provides additional insight into the composition of $\widehat{\boldsymbol{Y}}$.

Proposition 3. For $\{a_\lambda(\mathbf{j})\}$ given in Proposition 2, define

$$\pi(k) = \sum_{\mathrm{card}(\mathbf{j})=k} a_\lambda(\mathbf{j}) \quad \text{for } k = 0, \cdots, p.$$

Then

$$\pi(k) = P(V_1 + \cdots + V_p = k)$$

where $V_1, \cdots, V_p$ are independent 0/1 random variables with $P(V_i = 1) = \mu_i/(\mu_i + \lambda)$ and $\mu_1, \cdots, \mu_p$ are the eigenvalues of $\mathcal{X}^T\mathcal{X}$.

Proof. Define

$$\varphi(t) = \prod_{i=1}^{p} \left\{ \frac{\lambda}{\mu_i + \lambda} + t\frac{\mu_i}{\mu_i + \lambda} \right\} = E\left(t^{V_1 + \cdots + V_p}\right)$$

to be the probability generating function of $V_1 + \cdots + V_p$; we need to show that

$$\sum_{k=0}^{p} t^k \pi(k) = \varphi(t).$$

To do so, we simply need to match the coefficients of t^k for $k = 0, \cdots, p$; for $k = 0$ and $k = p$, this holds by inspection. Noting that $|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|$ are principal minors of $\mathcal{X}^T\mathcal{X}$, we have for $k = 1, \cdots, p-1$

$$\begin{aligned} \sum_{\mathrm{card}(\mathbf{j})=k} |\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})| &= \text{coefficient of } \lambda^{p-k} \text{ in } |\mathcal{X}^T\mathcal{X} + \lambda I| \\ &= \sum_{j_1 < \cdots < j_k} (\mu_{j_1}\mu_{j_2} \times \cdots \times \mu_{j_k}) \end{aligned}$$

and so

$$t^k\pi(k) = t^k \frac{\lambda^{p-k}}{(\mu_1 + \lambda) \times \cdots \times (\mu_p + \lambda)} \sum_{j_1 < \cdots < j_k} (\mu_{j_1} \times \cdots \times \mu_{j_k}).$$

The conclusion follows by noting that the coefficient of t^k in $\varphi(t)$ is

$$\frac{\lambda^{p-k}}{(\mu_1+\lambda)\times\cdots\times(\mu_p+\lambda)}\sum_{j_1<\cdots<j_k}(\mu_{j_1}\times\cdots\times\mu_{j_k}).$$

Given the eigenvalues $\mu_1,\cdots,\mu_p$ of $\mathcal{X}^T\mathcal{X}$, the distribution of $N = V_1+\cdots+V_p$ can computed using a discrete Fourier transform (DFT). The DFT of $\{\pi(k)\}$ is

$$\begin{aligned}\widehat{\pi}(s) &= \sum_{k=0}^{p}\pi(k)\exp\left(-2\pi\iota\frac{ks}{p+1}\right)\\ &= \prod_{j=1}^{p}\left\{\frac{\lambda}{\mu_j+\lambda}+\frac{\mu_j}{\mu_j+\lambda}\exp\left(-2\pi\iota\frac{s}{p+1}\right)\right\}\quad\text{for } s=0,\cdots,p\end{aligned}$$

(where $\iota=\sqrt{-1}$) and so applying the inverse DFT

$$\pi(k)=\frac{1}{p+1}\sum_{s=0}^{p}\widehat{\pi}(s)\exp\left(2\pi\iota\frac{ks}{p+1}\right).$$

Note that $\pi(k)$ is necessarily real-valued although numerically $\pi(k)$ as evaluated above will typically contain a small imaginary component due to round-off error; this is resolved by taking the real part of the inverse DFT.

For ridge regression, the vector of fitted values is $\widehat{\boldsymbol{y}} = A_\lambda\boldsymbol{y}$ where $A_\lambda = \mathcal{X}(\mathcal{X}^T\mathcal{X}+\lambda I)^{-1}\mathcal{X}^T$. The distribution of N over the integers $0,\cdots,p$ depends on the non-zero eigenvalues of A_λ and as such, could be viewed as a sort of spectrum of the ridge regression estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ (dependent on the spectrum of A_λ). In the next section, we will show that this simple observation generalizes if we replace A_λ in the definition of $\widehat{\boldsymbol{y}}$ by a matrix with real-valued eigenvalues lying in the interval $[0,1]$. This spectrum (distribution of N) might be viewed as analogous to the notion of timbre (or tone colour) in music whereby the frequency spectrum (as well as other characteristics) of a given note varies across different musical instruments (Samson *et al.*, 1997).

In a very crude sense, the distribution of N reflects the bias-variance tradeoff in ridge regression. Specifically, if $E(N)$ is close to p then the bias in $\widehat{\boldsymbol{\beta}}(\lambda)$ is typically smaller while the variance of $\widehat{\boldsymbol{\beta}}(\lambda)$ should decrease as $\mathrm{Var}(N)$ increases as a consequence of greater model averaging. In general, the distribution of N should provide more information about $\widehat{\boldsymbol{y}}$ than is provided by $E(N)$ alone. The following example illustrates how ridge regression does the "right thing" by averaging a broader range of models when the design has a high degree of multicollinearity.

Example 4. As mentioned earlier, variance inflation factors (VIFs) are used to assess the effects of multicollinearity on the variance of an OLS estimator of a parameter – they measure the inflation of the variance of the estimator relative to the variance in the case

where the predictors are uncorrelated. In the case where $\mathcal{X}^T\mathcal{X}$ is a correlation matrix (and non-singular) then the VIFs are simply the diagonal elements of $(\mathcal{X}^T\mathcal{X})^{-1}$.

Suppose that $\mathcal{X}^T\mathcal{X}$ is a correlation matrix with eigenvalues $\mu_1, \cdots, \mu_p$ where $\mu_1, \cdots, \mu_r$ are positive. Then

$$\begin{aligned} E(N) = E(N_\lambda) &= \sum_{j=1}^{p} \frac{\mu_j}{\mu_j + \lambda} \\ \mathrm{Var}(N) = \mathrm{Var}(N_\lambda) &= \sum_{j=1}^{p} \frac{\mu_j}{\mu_j + \lambda} - \sum_{j=1}^{p} \frac{\mu_j^2}{(\mu_j + \lambda)^2}. \end{aligned}$$

Differentiating with respect to λ, we have

$$\begin{aligned} \frac{d}{d\lambda} E(N_\lambda) &= -\sum_{j=1}^{p} \frac{\mu_j}{(\mu_j + \lambda)^2} \\ &\to -\sum_{j=1}^{r} \mu_j^{-1} \text{ as } \lambda \to 0 \\ \frac{d}{d\lambda} \mathrm{Var}(N_\lambda) &= 2\sum_{j=1}^{p} \frac{\mu_j^2}{(\mu_j + \lambda)^3} - \sum_{j=1}^{p} \frac{\mu_j}{(\mu_j + \lambda)^2} \\ &\to \sum_{j=1}^{r} \mu_j^{-1} \text{ as } \lambda \to 0. \end{aligned}$$

When $\mathcal{X}^T\mathcal{X}$ is non-singular then $\mu_1^{-1} + \cdots + \mu_p^{-1} = \mathrm{tr}((\mathcal{X}^T\mathcal{X})^{-1})$; thus for λ close to 0, $E(N_\lambda)$ is decreased by factor proportional to the sum of the VIFs for the p predictors while $\mathrm{Var}(N_\lambda)$ is increased by the same factor.

The form of the derivatives above suggests that N_λ has a Poisson-like distribution for small values of λ. Indeed if we assume that r (the number of positive eigenvalues) tends to infinity and λ tends to 0 so that

$$\lambda \sum_{j=1}^{r} \mu_j^{-1} \to \kappa > 0 \text{ and } \max_{1 \le j \le r} \frac{\lambda}{\mu_j} \to 0$$

then $r - N_\lambda \xrightarrow{d} \mathrm{Poisson}(\kappa)$. A similar result holds for large λ: If r and λ tend to infinity so that

$$\sum_{j=1}^{r} \frac{\mu_j}{\mu_j + \lambda} \to \kappa > 0 \text{ and } \max_{1 \le j \le r} \frac{\mu_j}{\mu_j + \lambda} \to 0$$

then $N_\lambda \xrightarrow{d} \mathrm{Poisson}(\kappa)$.

The results of Propositions 2 and 3 can be extended to negative values of λ provided that $\mathcal{X}^T\mathcal{X} + \lambda I$ is non-singular; we lose the probabilistic interpretation of $\widehat{\boldsymbol{\beta}}(\lambda)$ as the expected value $E_\lambda[\widehat{\boldsymbol{\beta}}_{\mathbf{S}}]$ for some random subset $\mathbf{S} \sim \mathcal{P}_\lambda$ but mathematically, we still have

$$\widehat{\boldsymbol{\beta}}_\lambda = \sum_{\mathbf{s}} |H_\lambda(\mathbf{s})| \widehat{\boldsymbol{\beta}}_{\mathbf{s}}$$

where now the off-diagonal blocks of H_λ in (3) are imaginary (although for any $\mathbf{s}$, $|H_\lambda(\mathbf{s})|$ is real-valued). Alternatively, we can redefine H_λ as follows:

$$H_\lambda = \begin{pmatrix} \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T & \lambda\mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \\ (\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T & \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} \end{pmatrix}.$$

H_λ is no longer a projection matrix (as it is not symmetric) but is still idempotent and we can use the results of Berman (1988) to prove that

$$\widehat{\boldsymbol{\beta}}(\lambda) = \sum_{\mathbf{s}} |H_\lambda(\mathbf{s})|\widehat{\boldsymbol{\beta}}_{\mathbf{s}}$$

where now $|H_\lambda(\mathbf{s})|$ need not lie in the interval $[0, 1]$ but

$$\sum_{\mathbf{s}} |H_\lambda(\mathbf{s})| = 1.$$

It can be shown that Propositions 2 and 3 still hold where now $\{a_\lambda(\mathbf{j})\}$ and $\{\pi(k)\}$ can take both negative values and values greater than 1 for certain subsets $\mathbf{j}$ and model sizes k, with both $\{a_\lambda(\mathbf{j})\}$ and $\{\pi(k)\}$ summing to 1.

In addition, we can also define the LoE of a predictor using the formula given in Proposition 1; depending on the value of $\lambda < 0$, the LoE of a given predictor could be greater than 1 or negative. However, if $\mu_1 \geq \cdots \geq \mu_r > 0$ are the positive eigenvalues of $\mathcal{X}^T\mathcal{X}$ and $\lambda > -\mu_r$ then $\text{LoE}_\lambda(j) \geq 0$ for all j; this follows from the fact that the matrix $I - \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}$ is non-negative definite under this condition on λ. Moreover, the "small λ" approximation for $\text{LoE}_\lambda(j)$ given in Example 1 also holds for negative values of λ.

The following example outlines how we might apply the result of Proposition 3 in the case where $\lambda < 0$.

Example 5. Kobak *et al.* (2020) demonstrate that under certain conditions, the optimal value of λ (in terms of MSE) could be negative. For example, suppose that for large $r = \text{rank}(\mathcal{X})$, we take $\lambda < 0$ such that

$$\sum_{j=1}^{r} \frac{\lambda}{\mu_j} = -\gamma < 0$$

(where $\mu_1, \cdots, \mu_r$ are the positive eigenvalues of $\mathcal{X}^T\mathcal{X}$) and

$$\max_{1 \leq j \leq r} \left| \frac{\lambda}{\mu_j} \right| \approx 0.$$

(In other words, λ is negative but very close to 0; for example, $\lambda = O(1/r)$ as $r \to \infty$.) Then

$$\pi(k) \approx \exp(\gamma)\frac{(-\gamma)^{r-k}}{(r-k)!}.$$

For large r, as $\gamma \to 0$, we have $\pi(r) = 1 + \gamma + o(\gamma)$ and $\pi(r-1) = -\gamma + o(\gamma)$ with $\pi(r-2), \pi(r-3), \cdots$ being $o(\gamma)$. This limiting case suggests that ridge regression with $\lambda < 0$ can potentially act as a jackknife-like bias reduction method (Quenouille, 1949, 1956; Schucany *et al.*, 1971). In particular, suppose that we want to estimate some real-valued parameter θ (for example $\theta = E(Y|\boldsymbol{x})$, a prediction of the response Y for some predictor value $\boldsymbol{x}$) and define $\widehat{\theta}_\lambda$ to be the ridge estimate for some $\lambda < 0$ where

$$\lambda = -\gamma \left\{ \sum_{j=1}^{r} \mu_j^{-1} \right\}^{-1}$$

for some small $\gamma > 0$. In this "small γ" scenario,

$$\widehat{\theta}_\lambda \approx (1+\gamma)\widehat{\theta}(r) - \gamma\,\widehat{\theta}(r-1).$$

where $\widehat{\theta}(r)$ is the least squares estimator of θ for an r dimensional model and $\widehat{\theta}(r-1)$ is a weighted average of least squares estimators of θ from $r-1$ dimensional models. Define $b(r)$ and $b(r-1)$ to be the respective biases of $\widehat{\theta}(r)$ and $\widehat{\theta}(r-1)$ and assume that either $0 < b(r) < b(r-1)$ or $0 > b(r) > b(r-1)$. Then the bias of $\widehat{\theta}_\lambda$ is (approximately)

$$E(\widehat{\theta}_\lambda) - \theta \approx b(r) + \gamma[b(r) - b(r-1)]$$

where for γ sufficiently small

$$|b(r) + \gamma[b(r) - b(r-1)]| < |b(r)|.$$

Similarly, we can approximate the MSE of $\widehat{\theta}_\lambda$:

$$\begin{aligned} \mathrm{MSE}(\widehat{\theta}_\gamma) \;\approx\; & (1+\gamma)^2 \mathrm{Var}[\widehat{\theta}(r)] + \gamma^2 \mathrm{Var}[\widehat{\theta}(r-1)] - 2\gamma(1+\gamma)\mathrm{Cov}[\widehat{\theta}(r), \widehat{\theta}(r-1)] \\ & + \{b(r) + \gamma[b(r) - b(r-1)]\}^2 . \end{aligned}$$

The derivative of the right hand side is negative at $\gamma = 0$, in which case the right hand is minimized at $\gamma > 0$ (that is, $\lambda < 0$), if

$$b(r)[b(r-1) - b(r)] > \mathrm{Var}[\widehat{\theta}(r)] - \mathrm{Cov}[\widehat{\theta}(r), \widehat{\theta}(r-1)].$$

Typically, $\mathrm{Var}[\widehat{\theta}(r)] - \mathrm{Cov}[\widehat{\theta}(r), \widehat{\theta}(r-1)]$ will be quite small but will increase as the variance of the noise increases. The bias terms $b(r)$ and $b(r-1)$ depend on the model error associated with r and $r-1$ dimensional models — if $|b(r)|$ is sufficiently large and $b(r-1) - b(r)$ has the same sign as $b(r)$ then taking $\lambda < 0$ will be optimal in terms of MSE.

The results of this paper focus on linear estimation where $\widehat{\boldsymbol{y}} = A\boldsymbol{y}$ for some fixed A. However, extensions to non-linear estimation are possible, for example, by looking a local linear approximations like $\widehat{\boldsymbol{y}} = A(\boldsymbol{y})\boldsymbol{y}$. The following example shows how the result of Proposition 3 might be applied to the LASSO (Tibshirani, 1996).

Example 6. The LASSO estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ minimizes

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where $\|\boldsymbol{\beta}\|_1$ is the L_1-norm of $\boldsymbol{\beta}$. As with ridge regression, we typically centre and scale the predictors so that $\mathcal{X}^T\mathcal{X}$ is the correlation matrix of the predictors.

Unlike ridge regression, the LASSO is not a linear estimation method although for each λ, we can write the estimate in a pseudo-linear form as follows:

$$\widehat{\boldsymbol{\beta}}(\lambda) = D^{1/2}(\lambda)\left([\mathcal{X}D^{1/2}(\lambda)]^T[\mathcal{X}D^{1/2}(\lambda)] + \frac{\lambda}{2}I\right)^{-1} D^{1/2}(\lambda)\mathcal{X}^T\boldsymbol{y}$$

where $D(\lambda)$ is a $p \times p$ diagonal matrix whose elements are the absolute values of the elements of $\widehat{\boldsymbol{\beta}}(\lambda)$. Thus if the j-th component of $\widehat{\boldsymbol{\beta}}(\lambda)$ equals 0 then the j-th column of $\mathcal{X}D^{1/2}(\lambda)$ will be a vector of zeroes and the rank of $\mathcal{X}D^{1/2}(\lambda)$ will be reduced (relative to the rank of $\mathcal{X}$) by the number of 0 components in $\widehat{\boldsymbol{\beta}}(\lambda)$. We can then apply Proposition 3 using the matrix $D^{1/2}(\lambda)\mathcal{X}^T\mathcal{X}D^{1/2}(\lambda)$ in place of $\mathcal{X}^T\mathcal{X}$ to determine the distribution of N (conditional on $\widehat{\boldsymbol{\beta}}(\lambda)$) for a given λ.

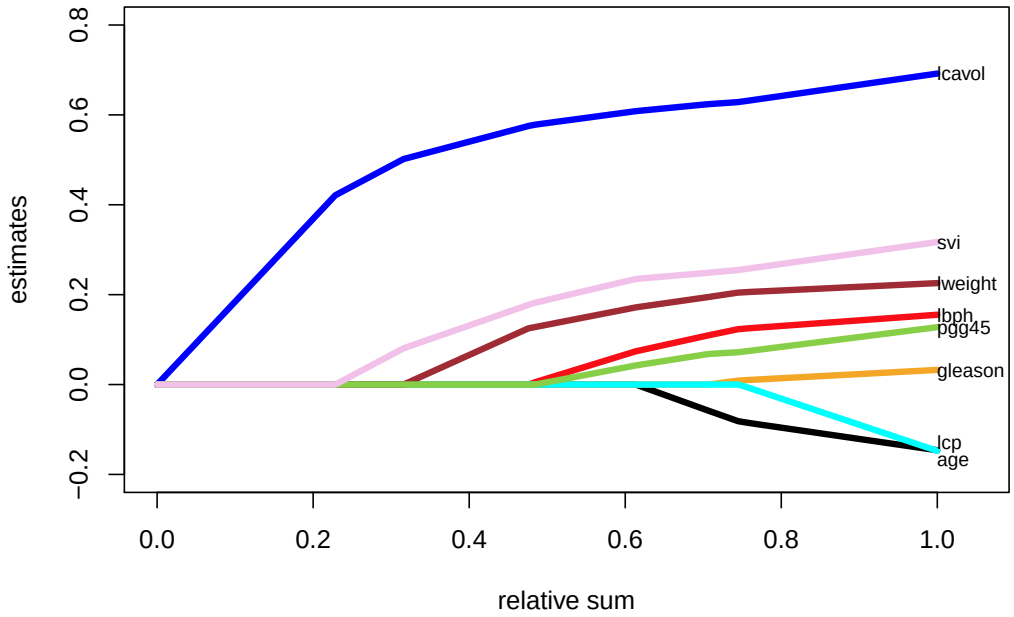

Figure 1: LASSO plot: Plot of the 8 estimates as a function of $\|\widehat{\boldsymbol{\beta}}(\lambda)\|_1/\|\widehat{\boldsymbol{\beta}}(0)\|_1$.

As an illustration, we consider the prostate cancer data described in Tibshirani (1996), where the relationship between the logarithm of prostate specific antigen (PSA) and 8 predictors (prognostic variables). Figure 1 is a plot of the elements of $\widehat{\boldsymbol{\beta}}(\lambda)$ versus the "relative

n	0	1	2	3	4	5	6	7	8
LASSO	0.000	0.000	0.000	0.000	0.006	0.053	0.233	0.446	0.262
Ridge	0.000	0.000	0.000	0.001	0.011	0.065	0.224	0.404	0.294

Table 1: Distribution of N for LASSO and ridge regression when $E(N) = 6.9$.

n	0	1	2	3	4	5	6	7	8
LASSO	0.003	0.054	0.237	0.400	0.255	0.051	0.000	0.000	0.000
Ridge	0.014	0.096	0.243	0.309	0.221	0.092	0.022	0.003	0.000

Table 2: Distribution of N for LASSO and ridge regression when $E(N) = 3.0$.

sum" $s(\lambda) = \|\widehat{\boldsymbol{\beta}}(\lambda)\|_1/\|\widehat{\boldsymbol{\beta}}(0)\|_1$. We consider $s(\lambda) = 0.9$ and $s(\lambda) = 0.6$; for $s(\lambda) = 0.9$, all elements of $\widehat{\boldsymbol{\beta}}(\lambda)$ are non-zero and $E(N) = 6.9$ while for $s(\lambda) = 0.6$, five are non-zero and $E(N) = 3.0$. Tables 1 and 2 show the distribution of N for the LASSO and ridge regression assuming $E(N) = 6.9$ and $E(N) = 3.0$, respectively. For $E(N) = 6.9$, the two distributions are similar while for $E(N) = 3.0$, the difference is more pronounced; in the latter case, the distribution of N for ridge regression is more dispersed while the distribution of N (conditional on $\widehat{\boldsymbol{\beta}}(\lambda)$) for the LASSO is necessarily concentrated on the integers 0, 1, $\cdots$, 5.

4 Smoothing matrices

Proposition 2 is remarkable in that it gives a decomposition of the ridge regression estimate in terms of OLS estimates for all possible 2^p subset models; Proposition 3 provides a means for obtaining a spectrum of the ridge regression estimate over subset models of size less than or equal to p. As noted above, Proposition 3 can also be expressed in terms of the matrix $A_\lambda = \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda)^{-1}\mathcal{X}^T$, which maps the vector of responses $\boldsymbol{y}$ to the vector of fitted values $\widehat{\boldsymbol{y}} = A_\lambda \boldsymbol{y}$. Proposition 2 implies that A_λ is a weighted average of projection matrices corresponding the 2^p subset OLS estimates while Proposition 3 effectively defines a probability distribution $\pi(k)$ on Grassmannians $\{\mathrm{Gr}(k,n) : k = 0, 1, \cdots, p\}$ where $\mathrm{Gr}(k,n)$ is the space parameterizing all k dimensional subspaces of R^n; $\mathrm{Gr}(k,n)$ is homeomorphic to the space of $n \times n$ projection matrices with trace k. Writing

$$A_\lambda = \sum_H p(H)H \quad \text{with} \quad \sum_H p(H) = 1$$

(where the sums above are over all projection matrices for the subset models) then

$$\pi(k) = \sum_{H:\mathrm{tr}(H)=k} p(H).$$

In this section, we will show that the results for ridge regression extend naturally to a more general class of matrices as described below.

Suppose that A is an $n \times n$ symmetric matrix with eigenvalues $\mu_1, \cdots, \mu_n \in [0,1]$; given a vector of responses $\boldsymbol{y}$, we define fitted values by $\widehat{\boldsymbol{y}} = A\boldsymbol{y}$. Such matrices are often used in non-parametric regression estimation to fit smooth functions to data and so we will refer to such matrices as smoothing matrices. More generally, we can relax the symmetry assumption by assuming that the singular values of A lie in $[0,1]$; if $A = U\Sigma V^T$ where U and V are orthogonal matrices and Σ is the diagonal matrix of singular values then defining $\boldsymbol{y}^* = UV^T\boldsymbol{y}$ (or $\boldsymbol{y} = VU^T\boldsymbol{y}^*$) so that $\widehat{\boldsymbol{y}} = U\Sigma U^T\boldsymbol{y}^*$. Note that if we assume the classical model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ then $\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{\mu}^* = UV^T\boldsymbol{\mu}, \sigma^2 I)$. Alternatively, we can symmetrize A, for example, by $A^* = I - [(I-A)^T(I-A)]^{1/2}$ as in Cohen (1966) or by $A^* = (A + A^T)/2$. We may also be able to approximate A by a symmetric matrix as in Milanfar (2013). With this in mind, we will assume henceforth that A is symmetric.

First of all, for a general smoothing matrix A, there is not necessarily a unique representation $A = \sum_H p(H)H$. Choi and Wu (1990) show that a smoothing matrix A can be expressed as $A = w_1H_1 + \cdots + w_mH_m$ for positive $w_1, \cdots, w_m$ summing to 1 and projection matrices $H_1, \cdots, H_m$; moreover, they show that $m \leq \lfloor \log_2(n) \rfloor + 2$ where the upper bound on m is sharp. Our goal (in the spirit of Proposition 2) is to find a maximal representation for A in terms of projection matrices. For example, if $A = (H_1 + \cdots + H_m)/m$ where $H_1, \cdots, H_m$ are all (say) one-dimensional projection matrices then it may be tempting to say that $\pi(k)$ (as defined in Proposition 3) will have $\pi(1) = 1$ with $\pi(k) = 0$ for $k \neq 1$; however as we will show subsequently, $\pi(k)$ will be determined by the eigenvalues of A. (Moreover in the case where $A = w_1H_1 + \cdots + w_mH_m$, the condition that $w_1 + \cdots + w_m = 1$ is not necessary; we merely need the eigenvalues of A to lie between 0 and 1.)

It is simple to show that Proposition 3 holds for smoothing matrices. Suppose that A has r non-zero eigenvalues $\mu_1, \cdots, \mu_r$ (with $\mu_{r+1} = \cdots = \mu_n = 0$). Then $A = \Gamma_+ D_+ \Gamma)+^T$ where D_+ is an $r \times r$ diagonal matrix consisting of the non-zero eigenvalues $\mu_1, \cdots, \mu_r$ and the columns of Γ_+ are the orthonormal eigenvectors corresponding to the non-zero eigenvalues. Thus we can write

$$\widehat{\boldsymbol{y}} = A\boldsymbol{y} = \Gamma_+\widehat{\boldsymbol{\beta}}$$

with $\widehat{\boldsymbol{\beta}} = D_+\Gamma_+^T\boldsymbol{y}$. Then $\widehat{\boldsymbol{\beta}}$ minimizes the generalized ridge regression objective function

$$\|\boldsymbol{y} - \Gamma_+\boldsymbol{\beta}\|^2 + \sum_{j=1}^{r} \frac{1-\mu_j}{\mu_j}\beta_j^2 = \|\boldsymbol{y} - \Gamma_+\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^T(D_+^{-1} - I)\boldsymbol{\beta}.$$

Thus the results of Propositions 2 and 3 can be extended *mutatis mutandis* to symmetric

smoothing matrices. Writing

$$\widehat{\boldsymbol{\beta}} = \sum_{\text{all } \mathbf{j}} a(\mathbf{j})\widehat{\boldsymbol{\beta}}(\mathbf{j}),$$

we have

$$a(\mathbf{j}) = \left\{\prod_{h \notin \mathbf{j}} (1-\mu_h)\right\}\left\{\prod_{h \in \mathbf{j}} \mu_h\right\}.$$

Thus

$$\begin{aligned}\pi(k) &= \sum_{\text{card}(\mathbf{j})=k} a(\mathbf{j}) \\ &= P(N = V_1 + \cdots + V_r = k)\end{aligned}$$

where as before $V_1, \cdots, V_r$ are independent random variables with $P(V_i = 1) = \mu_i$ and $P(V_i = 0) = 1 - \mu_i$. From this, it follows that if $\pi(k) = 1$ for some k then A must be a projection matrix (with rank k) since $N = k$ with probability 1 if, and only if, k of $V_1, \cdots, V_r$ equal 1 and $r - k$ equal 0 with probability 1.

4.1 Ensemble estimation

Ensemble estimation combines simple (low dimensional) models with a goal of reducing prediction error. Its effectiveness in practice can be explained roughly by the premise that any bias in prediction from low dimensional models is offset by a reduction in variance due to averaging. For example, if $\widehat{\boldsymbol{y}}_h = A_h\boldsymbol{y}$ for $(h = 1, \cdots, m)$ are predictions from m models then we can define

$$\widehat{\boldsymbol{y}}_0 = \left(\sum_{h=1}^{m} w_h A_h\right)\boldsymbol{y} = A_0\boldsymbol{y}$$

for some weights $w_1, \cdots, w_m$. It is also worth noting that ensemble estimation may be computationally advantageous if the computational cost of the low dimensional models is relatively small compared to that of higher dimensional models.

Given that ridge regression can be considered as an emsemble estimation method, a natural question to ask is to what extent we can approximate an ensemble method by a simple ridge regression with tuning parameter λ; in other words, is

$$\sum_{h=1}^{m} w_h A_h \approx \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T$$

for some λ? (For example, if the approximation is sufficiently good, we could use the value of λ to approximate the LoEs of each predictor.) While this question is beyond the scope of this paper, Example 7 below suggests that such an approximate equivalence may not be far-fetched. Moreover, recent work by LeJeune *et al.* (2019) suggests that the connection between this type of ensemble estimation and ridge regression may have a more solid theoretical foundation.

A simple (somewhat trivial) example where there is an exact equivalence between ensemble estimation and ridge regression occurs when $\mathcal{X}^T\mathcal{X} = I$ where we take

$$A_h = \frac{1}{1+\lambda_0}\mathcal{X}_h\mathcal{X}_h^T$$

$h = 1, \cdots, m = \binom{p}{k}$ subsets of $k < p$ predictors; in other words, we are doing ridge regression with parameter λ_0 for each subset of predictors. Then

$$\frac{1}{m}\sum_{h=1}^{m} A_h = \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T$$

where

$$\lambda = \frac{p}{k}(1+\lambda_0) - 1.$$

More generally, we might expect this type of ensemble estimation to be a good approximation to ridge regression (or vice versa) when the p predictors are weakly correlated, for example, if the variability of the non-zero eigenvalues of the predictor correlation matrix is small.

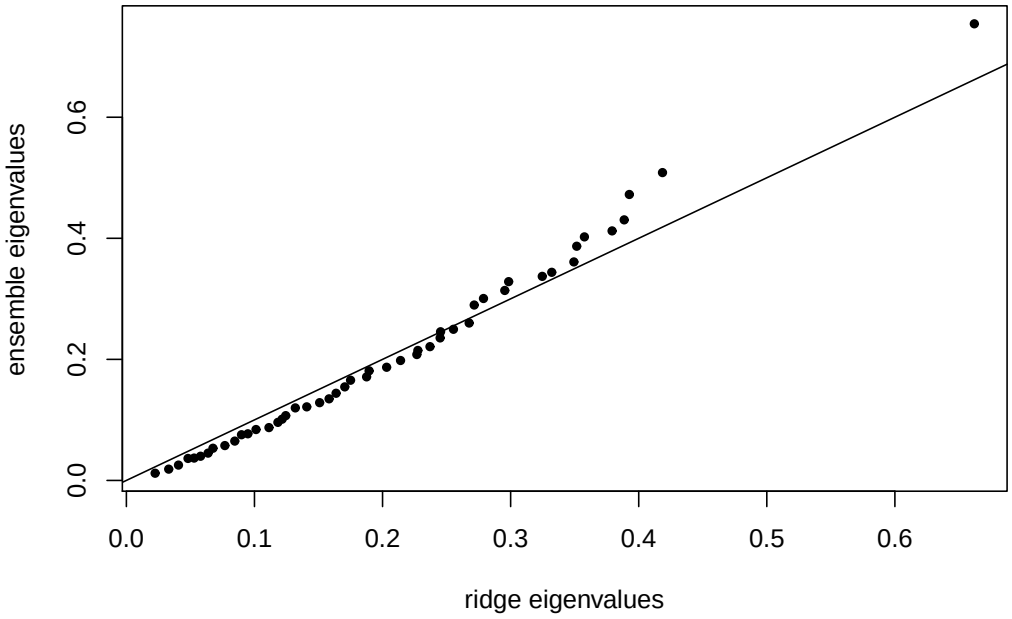

Figure 2: Eigenvalues of A for the ensemble method in Example 7 versus eigenvalues for ridge regression.

Example 7. We take $p = 100$ and $n = 50$, and define $\mathcal{X}$ so that its columns have a fairly complex correlation structure with pairwise correlations ranging from -0.46 to 0.81. The values of $\mathrm{LoE}_0(j)$ (for $j = 1, \cdots, 100$) as defined in Example 1 range from 0.19 to 0.65; the

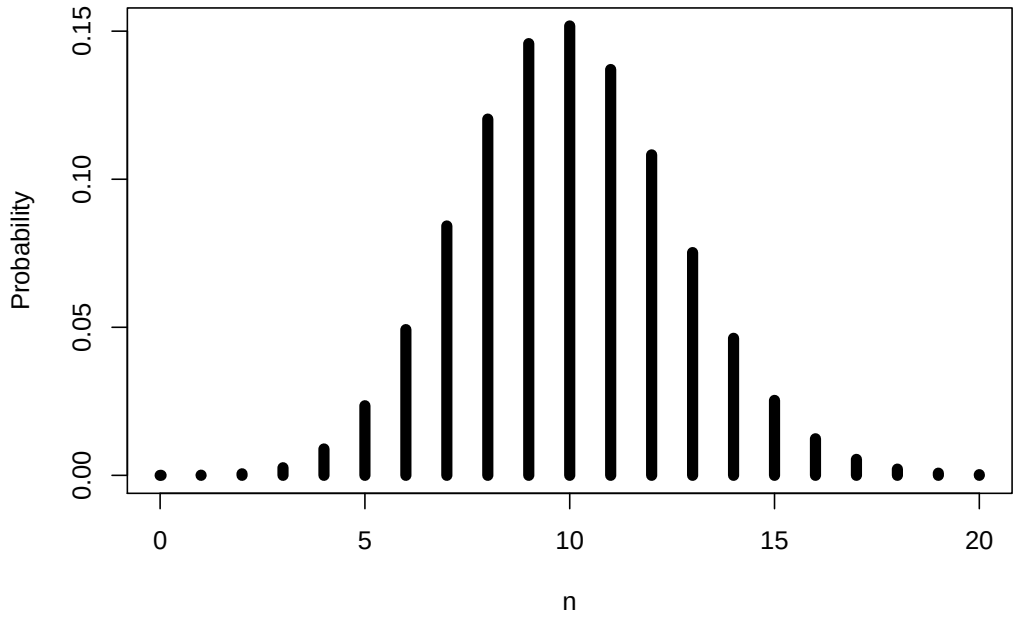

Figure 3: Distribution of N for the ensemble method in Example 7.

ad hoc bounds for $\mathrm{LoE}_0(j)$ given in Example 1 based on uniformly distributed orthogonal matrices are 0.33 and 0.72 with $\mathrm{LoE}_0(j)$ less than lower bound for two predictors. We then compute $\widehat{\boldsymbol{y}}$ by taking an average of 20 OLS estimates, each of which is based on a random selection of 10 (of 100) predictors; the matrix A is an average of $m = 20$ (randomly sampled) projection matrices whose traces are equal to 10. As a point of comparison, we also consider a ridge regression estimate with $\lambda = 6.77$ chosen so that $E(N_\lambda) = 10$. Figure 2 shows a plot of the 49 non-zero eigenvalues for the two methods while Figure 3 gives the distribution of N for the ensemble method. The distributions of eigenvalues are "close" with the ensemble method having more eigenvalues close to 0 and to 1 than ridge regression. Thus the variance of N for the ensemble method is somewhat smaller than it is for ridge regression (6.74 versus 7.16).

What happens if we increase the number m of estimates used to compute the ensemble estimate? As indicated in Table 3, the variance of N increases with m although it is still smaller than the variance for ridge regression. Table 3 also gives the L_2-norm of the difference between the matrices $A_r = \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)\mathcal{X}^T$ (ridge regression) and $A_e = m^{-1}\sum_{\ell=1}^{m} H_\ell$ (ensemble estimation) for different values of m; it shows that as we increase m, A_e is approximated better by A_r.

m	20	50	100	500	1000
$\mathrm{Var}(N)$	6.74	6.92	7.04	7.09	7.11
$\|A_r - A_e\|_2$	0.209	0.155	0.095	0.054	0.048

Table 3: $\mathrm{Var}(N)$ and $\|A_r - A_e\|_2$ for ensemble estimation as a function of m. For reference, $\|A_r\|_2 = 0.662$.

4.2 Approximating the distribution of N

So far, we have implicitly assumed that the number of predictors p was sufficiently small that the eigenvalues of $\mathcal{X}^T\mathcal{X}$ could be computed exactly with little difficulty. In the more general setting (where often A is not explicitly computed), computing the eigenvalues of A can be more difficult, in which case we would need to resort to approximations (Normal or Poisson) or Monte Carlo evaluation of the spectrum of A. Typically, we will know that number of eigenvalues equal to 1; for example, if A is used in non-parametric function estimation, we may know that A will preserve low degree polynomials up to degree k, in which case we will have $k+1$ eigenvalues equal to 1 and so $N = k + 1 + V_{k+2} + \cdots + V_r$.

Hutchinson's method (Hutchinson, 1990; Skilling, 1989) can be used to estimate both $\mathrm{tr}(A)$ and $\mathrm{tr}(A^2)$, which can be used to derive simple approximations of the distribution of N, based on either the Normal or Poisson distribution. If $\boldsymbol{U}$ is a random vector with mean $\mathbf{0}$ and covariance matrix I then $\mathrm{tr}(A) = E[\boldsymbol{U}^T A \boldsymbol{U}]$ and so we can estimate $\mathrm{tr}(A)$ and $\mathrm{tr}(A^2)$ by

$$\begin{aligned} \widehat{\mathrm{tr}}(A) &= \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{U}_j^T A \boldsymbol{U}_j \\ \widehat{\mathrm{tr}}(A^2) &= \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{U}_j^T A^2 \boldsymbol{U}_j \end{aligned}$$

where $\boldsymbol{U}_1, \cdots, \boldsymbol{U}_m$ are independent random vectors with mean $\mathbf{0}$ and covariance matrix I; the variance of the estimates are minimized if the components of $\{\boldsymbol{U}_j\}$ are independent Rademacher random variables taking values ± 1 each with probability 1.

Example 8. Consider $\widehat{\boldsymbol{y}} = A\boldsymbol{y}$ where

$$A^{-1} = \begin{pmatrix} 1+6\lambda & -4\lambda & \lambda & 0 & 0 & \cdots & 0 & \lambda & -4\lambda \\ -4\lambda & 1+6\lambda & -4\lambda & \lambda & 0 & \cdots & 0 & 0 & \lambda \\ \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & \cdots & 0 & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ -4\lambda & \lambda & 0 & 0 & 0 & \cdots & \lambda & -4\lambda & 1+6\lambda \end{pmatrix}$$

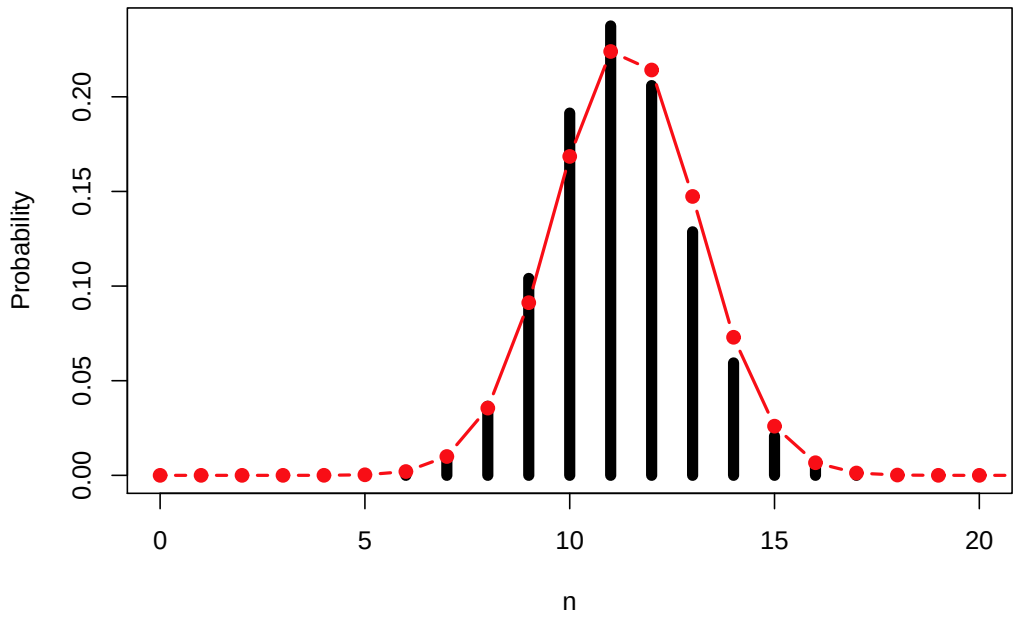

Figure 4: Exact distribution of N (black bars) in Example 8 for $\lambda = 1600$ and $n = 200$ compared to a Normal approximation (red) with mean and variance determined by Hutchinson's method.

This particular method may be appropriate if $y_i = g(i/n) + \varepsilon_i$ where g is a smooth function with $g(0) = g(1)$. The matrix A^{-1} is circulant and has eigenvalues

$$\mu_i^{-1} = 1 + 6\lambda - 8\lambda \cos\left(2\pi \frac{i}{n}\right) + 2\lambda \cos\left(4\pi \frac{i}{n}\right) \quad \text{for } i = 1, \cdots, n.$$

Figure 4 shows that exact distribution of N (for $\lambda = 1600$ and $n = 200$) and a Normal approximation with the mean and variance of N computed using Hutchinson's method using $m = 50$ replications; the approximation, while not perfect, gives us a good sense of the distribution of N.

5 Twicing and boosting

Twicing was introduced by Tukey (1977) as a means of refining the estimate $\widehat{\boldsymbol{y}}_1 = A\boldsymbol{y}$ by applying A to the residuals $\boldsymbol{y} - \widehat{\boldsymbol{y}}_1 = (I - A)\boldsymbol{y}$ and adding the result to $\widehat{\boldsymbol{y}}_1$:

$$\widehat{\boldsymbol{y}}_2 = \widehat{\boldsymbol{y}}_1 + A(I - A)\boldsymbol{y} = (2A - A^2)\boldsymbol{y}.$$

This procedure can be applied iteratively giving

$$\widehat{\boldsymbol{y}}_k = [I - (I - A)^k]\boldsymbol{y} = A_k\boldsymbol{y}$$

for $k = 1, 2, \cdots$. The eigenvalues of A_k are $1 - (1 - \mu_i)^k$ $(i = 1, \cdots, n)$ and it follows that as $k \to \infty$, A_k converges to a projection matrix to the spaced spanned by the eigenvectors of A with eigenvalues $\mu_i \neq 0$.

Related to twicing is the notion of boosting (Freund and Shapire, 1995), which is used in machine learning as a means of combining "weak learners" for prediction or classification, while reducing the possibility of overfitting. Bühlmann and Yu (2003) consider some the statistical aspects of boosting.

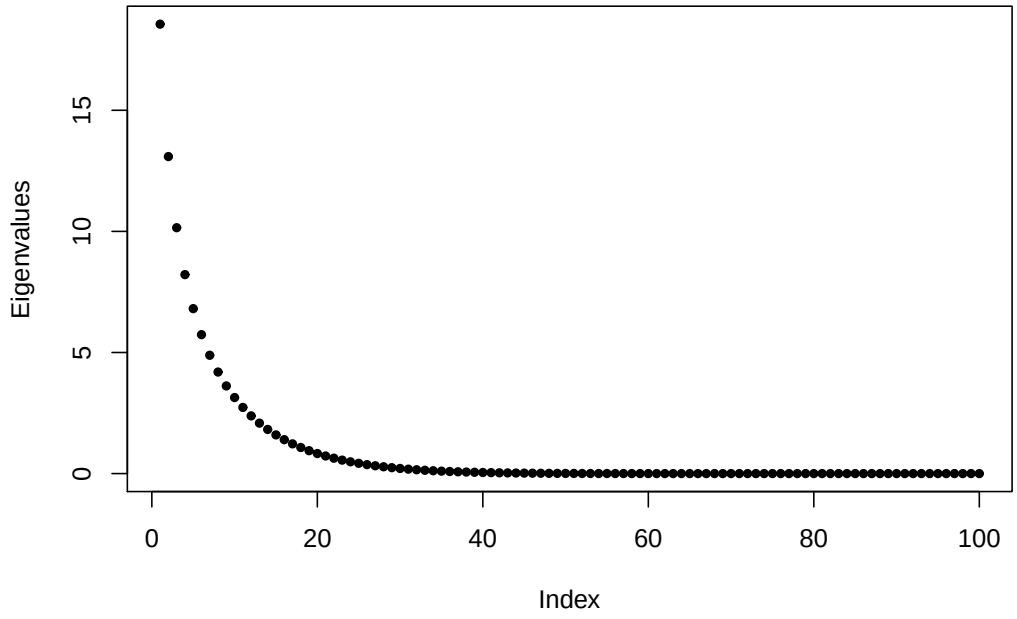

Figure 5: Eigenvalues of $\mathcal{X}^T\mathcal{X}$ in Example 9.

Example 9. In the case of ridge regression, $A = A_\lambda = \mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T$ and assume that $\mu_1, \cdots, \mu_p$ are the eigenvalues of $\mathcal{X}^T\mathcal{X}$ so that the eigenvalues of $A_{k,\lambda} = I - (I - A_\lambda)^k$ are $1 - [\lambda/(\mu_j + \lambda)]^k$ for $j = 1, \cdots, p$.

Now suppose that both λ and k tend to infinity so that $k/\lambda \to \tau > 0$; the eigenvalues of $A_{k,\lambda}$ tend to $1 - \exp(-\tau\mu_j)$ for $j = 1, \cdots, p$. We can then compare the distributions of N for ridge regression and "boosted" ridge regression assuming that the degrees of freedom $E(N) =$ some specified r; in other words, we choose λ and τ so that

$$\sum_{j=1}^{p} \frac{\mu_j}{\mu_j + \lambda} = r = \sum_{j=1}^{p} \{1 - \exp(-\tau\mu_j)\}.$$

To illustrate, we take $\mu_1, \cdots, \mu_n$ to be the $1/101, 2/101, \cdots, 100/101$ quantiles of a Gamma distribution with shape parameter 0.1, normalized to sum to 100; these are shown in Figure

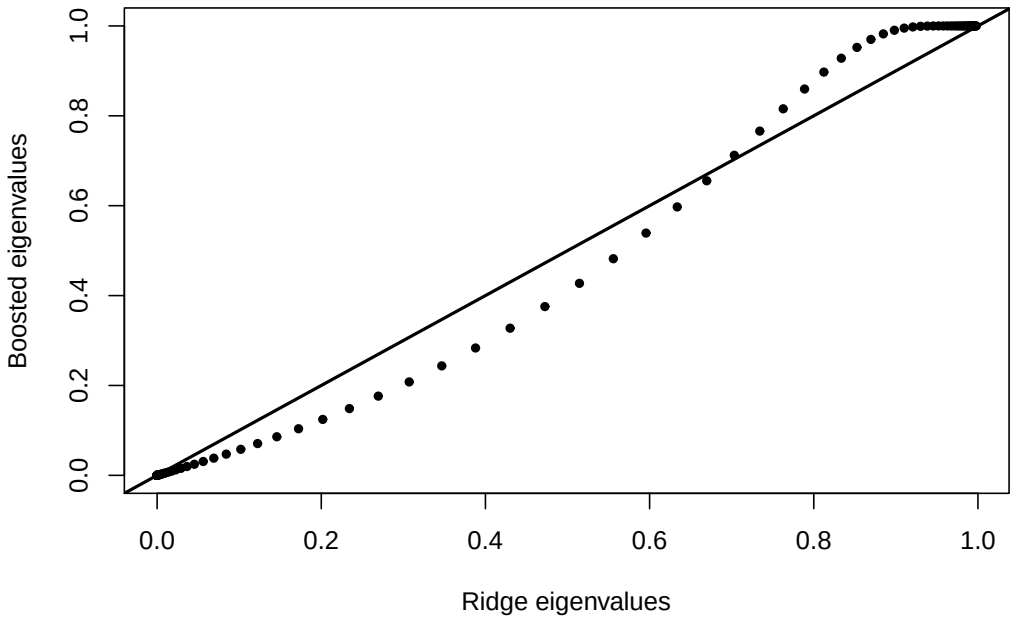

Figure 6: Plot of boosted eigenvalues versus eigenvalues for ridge regression.

5. Taking $E(N) = 40$, we get $\tau = 14.37$ and $\lambda = 0.03659$. (For example, if we take $k = 100$ then at each step, we would take the "base" value of $\lambda = 100/14.37 = 6.96$ with degrees of freedom $\sum_j \mu_j/(6.96 + \mu_j) = 7.49$; if $k = 1000$, the degrees of freedom at each step is 1.29.) Figure 6 gives plot of $\{1 - \exp(-\tau\mu_j\}$ versus $\{\mu_j/(\mu_j + \lambda)\}$; boosting increases the contribution of the larger eigenvalues while shrinking the contribution of the smaller eigevalues – in essence, this might be thought of as a sort of "soft" principal component (PC) regression where the contributions of smaller PCs are downweighted rather than eliminated. Figure 7 shows the distributions of N for ridge regression and boosted ridge regression. Note that boosting reduces the variability of N; $\mathrm{Var}(N) = 5.92$ for ridge regression compared to $\mathrm{Var}(N) = 4.03$ for boosted ridge regression.

6 Miscellanea

6.1 Continuous-time gradient descent

Example 9 suggests a connection between boosted ridge regression with large λ and gradient descent for OLS. When λ is large then

$$(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1} = \frac{1}{\lambda}I - \frac{1}{\lambda^2}\mathcal{X}^T\mathcal{X} + \cdots$$

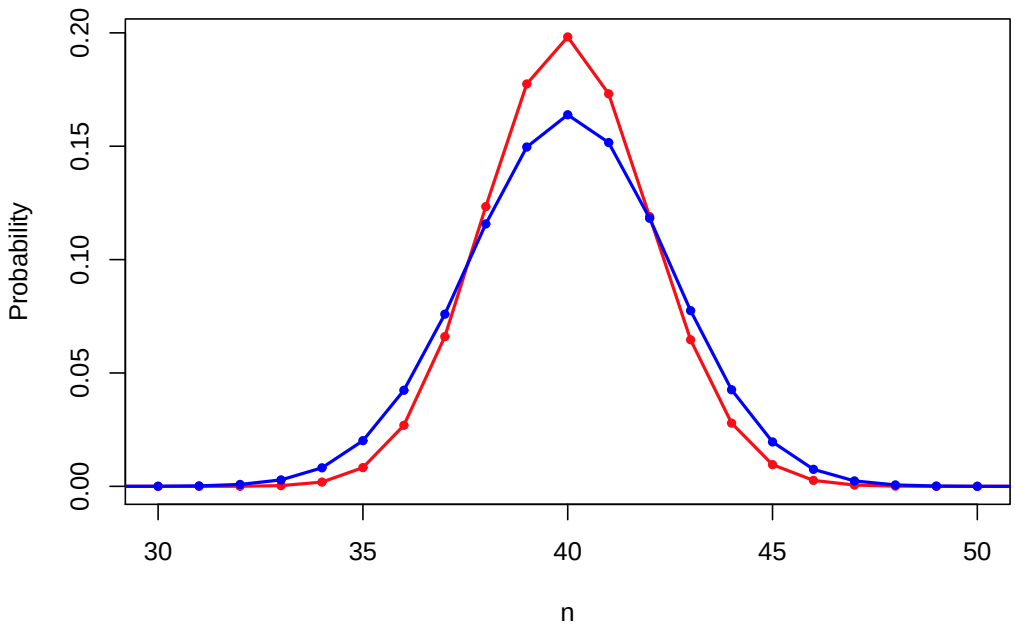

Figure 7: Distribution of N for ridge regression (blue) and boosted ridge regression (red).

so that

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k - \frac{1}{\lambda}(\mathcal{X}^T\mathcal{X}\widehat{\boldsymbol{\beta}}_k - \mathcal{X}^T\boldsymbol{y}) + o(\lambda^{-1}),$$

which, neglecting the $o(\lambda^{-1})$ term, is a gradient descent iteration for OLS.

By letting $\lambda \to \infty$, we can interpret the boosting process as approximating gradient descent for OLS in continuous-time resulting in $\widehat{\boldsymbol{\beta}}(t)$ satisfying the differential equation (or gradient flow)

$$\frac{d}{dt}\widehat{\boldsymbol{\beta}}(t) = -\mathcal{X}^T\mathcal{X}\widehat{\boldsymbol{\beta}}(t) + \mathcal{X}^T\boldsymbol{y}$$

(Skouras *et al.*, 1994; Ali *et al.*, 2018) whose solution given an initial estimate $\widehat{\boldsymbol{\beta}}(0)$ is

$$\widehat{\boldsymbol{\beta}}(t) = \widehat{\boldsymbol{\beta}}(0) + g(\mathcal{X}^T\mathcal{X}; t)\mathcal{X}^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}(0))$$

where (writing $\mathcal{X}^T\mathcal{X} = \Gamma D\Gamma^T$) $g(\mathcal{X}^T\mathcal{X}; t) = \Gamma g(D; t)\Gamma^T$ with g acting on the eigenvalues (diagonals of D) as follows:

$$g(\mu; t) = \begin{cases} 0 & \text{if } \mu = 0 \\ (1 - \exp(-t\mu))/\mu & \text{if } \mu > 0. \end{cases}$$

As $t \to \infty$, it is easy to see that $\widehat{\boldsymbol{\beta}}(t)$ converges to an OLS estimate.

Now assume that p is large and define $N(t)$ to the random variable in Proposition 3 corresponding to $\widehat{\boldsymbol{\beta}}(t)$. For simplicity, we will assume that $\widehat{\boldsymbol{\beta}}(0) = \mathbf{0}$ and so for a given t,

the fitted values $\widehat{\boldsymbol{y}}(t) = A(t)\boldsymbol{y}$ where $A(t) = \mathcal{X}g(\mathcal{X}^T\mathcal{X};t)\mathcal{X}^T$; as $t \to \infty$, $A(t)$ converges to a projection matrix with trace $r = \text{rank}(\mathcal{X})$ and the non-zero eigenvalues of $A(t)$ are $1 - \exp(-t\mu_j)$ for $j = 1, \cdots, r$ where $\mu_1, \cdots, \mu_r$ are the non-zero eigenvalues of $\mathcal{X}^T\mathcal{X}$.

We will consider two asymptotic scenarios for the non-zero eigenvalues of $A(t)$, taking r, the number of non-zero eigenvalues of $\mathcal{X}^T\mathcal{X}$, to infinity. The first scenario assumes that the distribution of the eigenvalues is well-behaved in the sense that

$$\sum_{j=1}^{r} \exp(-t\mu_j)(1 - \exp(-t\mu_j)) \to \infty \tag{5}$$

as $r \to \infty$; (5) holds if the empirical distribution of the eigenvalues converges to a probability distribution ν on the positive real line. The second scenario assumes that the distribution of the eigenvalues is heavy-tailed in the sense that for some sequence of constants $\{a_r\}$, we have

$$\sum_{j=1}^{r} I(a_r^{-1}\mu_j \in \cdot) \xrightarrow{v} M(\cdot) \text{ with } \int \mu\, M(d\mu) < \infty \tag{6}$$

for some point measure M on $(0, \infty)$ where "$\xrightarrow{v}$" denotes vague convergence of measures; see Kallenberg (1983) and Resnick(2013) for details regarding vague convergence. The limiting point measure M can be represented by a finite or countably infinite set of positive points $v_1, v_2, \cdots$. This latter scenario might be relevant in cases where $\mu_1, \cdots, \mu_r$ are widely dispersed so that (say) $\mu_1, \cdots, \mu_d$ are very large (possibly tending to infinity, in which case $a_r \to \infty$) with the remaining eigenvalues being significantly smaller.

Under (5), $N(t)$ will be approximately Normal with mean $\sum_{j=1}^{r}(1 - \exp(-t\mu_j))$ and variance $\sum_{j=1}^{r} \exp(-t\mu_j)(1 - \exp(-t\mu_j))$ while under (6), $N(t/a_r)$ converges in distribution (as $r \to \infty$) to a random variable

$$N_0(t) = \sum_{j=1}^{\infty} Y_j(t)$$

where $\{Y_j(t)\}$ are random variables taking values 0 and 1 with $P(Y_j(t) = 1) = 1 - \exp(-v_j t)$.

6.2 Approximating PC regression

In recent years, there has been some interest in iterative methods for approximating PC regression whereby we replace the $(n \times p)$ design matrix $\mathcal{X}$ by $\mathcal{X}\Gamma_\lambda$ where Γ_λ is an $p \times r$ matrix whose columns are the orthogonal eigenvectors corresponding to the r eigenvalues of $\mathcal{X}^T\mathcal{X}$ exceeding some threshold λ. In practice, such approximations essentially boil down to finding polynomial approximations to the projection matrix H_λ onto the column space of $\mathcal{X}\Gamma_\lambda$; specifically, we want to find a polynomial ψ_m and an $n \times n$ matrix A_λ so that

$$\psi_m(A_\lambda) = \sum_{k=1}^{m} \alpha_k A_\lambda^k \approx H_\lambda.$$

Ridge regression provides a starting point for these approximations. If $\mu_1, \cdots, \mu_p$ are the eigenvalues of $\mathcal{X}^T\mathcal{X}$ then $\mu_h/(\mu_h + \lambda) > 1/2$ if $\mu_h > \lambda$. We can then find a sequence of polynomials $\{\psi_m(x)\}$ such that

$$\psi_m\left(\frac{\mu}{\mu+\lambda}\right) \to \begin{cases} 1 & \text{if } \mu > \lambda \\ 0 & \text{if } \mu < \lambda \end{cases}$$

as $m \to \infty$. Then m sufficiently large,

$$\psi_m\left(\mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T\right)\boldsymbol{y} \tag{7}$$

can be used as an approximation for PC regression. See, for example, Frosting *et al.* (2016), Allen-Zhu and Li (2017), and Farnham *et al.* (2019) for implementations of this idea. The estimateof $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_m$ can be written as

$$\widehat{\boldsymbol{\beta}}_m = (\mathcal{X}^T\mathcal{X} + \Gamma W_m \Gamma^T)^{-1} X^T \boldsymbol{y}$$

where W_m is a diagonal matrix with elements

$$w_h = \mu_h \left\{ \psi_m^{-1}\left(\frac{\mu_h}{\mu_h+\lambda}\right) - 1 \right\} \quad \text{if } \mu_h > 0$$

with $w_h = \lambda$ if $\mu_h = 0$. Note that when $\mu_h > 0$, $w_h \to 0$ if $\mu_h > \lambda$ and $w_h \to \infty$ if $\mu_h < \lambda$ (as $m \to \infty$) as in Example 3.

Bernstein polynomials (Bernstein, 1912) provide one possible approach to constructing $\{\psi_m\}$ in the approximation (7). If $g(x)$ is a continuous function on $[0, 1]$, we can approximate it using the values $\{g(j/m) : j = 0, \cdots, m\}$ as follows:

$$\widetilde{g}_m(x) = \sum_{j=0}^{m} g(j/m)\binom{m}{j} x^j (1-x)^{m-j}$$

where $\{\binom{m}{j} x^j (1-x)^{m-j}\}$ are the Bernstein polynomials. If g is continuous then $\widetilde{g}_m$ converges uniformly on $[0, 1]$ to g as $m \to \infty$. In our context, we want $\psi_m(x)$ to approximate $\psi(x)$, which is 0 or 1 depending on whether $x < 1/2$ or $x > 1/2$. Taking m to be even (for convenience), we can define

$$\begin{aligned} \psi_m(x) &= \sum_{j=m/2}^{m} \binom{m}{j} x^j (1-x)^{m-j} \\ &= \sum_{j=m/2}^{m} \sum_{k=j}^{m} \binom{m}{k}\binom{k}{j} (-1)^{k-j} x^k \\ &= \sum_{k=m/2}^{m} \sum_{j=m/2}^{k} \binom{m}{k}\binom{k}{j} (-1)^{k-j} x^k. \end{aligned}$$

When m is reasonably large, $\psi_m(x)$ can be approximated in terms of the standard Normal distribution Φ as follows:

$$\psi_m(x) \approx \Phi\left(\frac{\sqrt{m}(x-1/2)}{\sqrt{x(1-x)}}\right).$$

Uniform convergence will not hold over $[0,1]$ although it will hold outside of a neighbourhood of $x = 1/2$. Moreover (and importantly so), the eigenvalues of $\psi_m(\mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T)$ will fall in the interval $[0,1]$ for all m . Using the eigenvalues in Example 9 and taking $\lambda = 0.0457$ we obtain $E(N) = 39.8$ and $\mathrm{Var}(N) = 1.08$ for $m = 40$ and $E(N) = 39.6$ and $\mathrm{Var}(N) = 0.68$ for $m = 100$.

The downside of the Bernstein polynomial approach described in the previous paragraph is its leisurely convergence rate; the example given there suggests that we would need to take m to be fairly large in order to approximate PC regression well. However, its spirit may be retained even if m is relatively small. It is worth remembering that PC regression is essentially an *ad hoc* procedure based on a bet that most of the predictive or explanatory power of the predictors in $\mathcal{X}$ will reside in the linear combinations (PCs) of predictors having largest variances or alternatively, that the predictive power of low variance PCs is minimal. Thus any procedure that amplifies the PCs with larger variances while attenuating the PCs with smaller variances essentially achieves the same goal as PC regression. ("Betting on sparsity", as coined in Hastie *et al.* (2001), is a similar principle that assumes that a response will be well-predicted by a small fraction of available predictors.) This view is somewhat controversial (Cox, 1968; Cook, 2007) although Artemiou and Li (2009) provide a theoretical framework to justify PC regression.

Appendix: Proof of Proposition 2.

Define $\mathbf{j} = \{j_1, \cdots, j_k\}$ to be a subset of $\{1, \cdots, p\}$ and

$$\mathbf{v_j} = n + \mathbf{j}^c = \{n + j : j \notin \{j_1, \cdots, j_k\}\}.$$

If $\mathbf{S} \cap \{n+1, \cdots, n+p\} = \mathbf{v_j}$ then the elemental estimate depends on the predictors in $\mathbf{j}$ with the parameter estimates for the predictors not in $\mathbf{j}$ set to 0.

Now define the event

$$\mathcal{A}_\mathbf{j} = \{\mathbf{S} \cap \{n+1, \cdots, n+p\} = \mathbf{v_j}\}$$

where $a_\lambda(\mathbf{j}) = P(\mathcal{A}_\mathbf{j})$. If $\mathbf{s} \cap \{n+1, \cdots, n+p\} = \mathbf{v_j}$, $X(\mathbf{s})$ can be written as an upper triangular block matrix where the lower diagonal is λ times the $(p-k) \times (p-k)$ identity matrix, which implies that $|X(\mathbf{s})|^2 = \lambda^{p-k}|\mathcal{X}_\mathbf{j}(\mathbf{s} \setminus \mathbf{v_j})|^2$ where $\mathcal{X}_\mathbf{j}(\mathbf{s} \setminus \mathbf{v_j})$ is the $k \times k$ sub-matrix

of $\mathcal{X}$ whose rows are in $\mathbf{s} \setminus \mathbf{v_j}$ (which is a subset of $\{1, \cdots, n\}$) and whose column indices are in $\mathbf{j}$.

Thus we have

$$\begin{aligned} P(\mathbf{S} = \mathbf{s}|\mathcal{A}_\mathbf{j}) &= \frac{|X(\mathbf{s})|^2}{\sum_\mathbf{u} |X(\mathbf{u})|^2} \\ &= \frac{\lambda^{p-k}|\mathcal{X}_\mathbf{j}(\mathbf{s} \setminus \mathbf{v_j})|^2}{\sum_\mathbf{u} \lambda^{p-k}|\mathcal{X}_\mathbf{j}(\mathbf{u} \setminus \mathbf{v_j})|^2} \\ &= \frac{|\mathcal{X}_\mathbf{j}(\mathbf{s} \setminus \mathbf{v_j})|^2}{\sum_\mathbf{u} |\mathcal{X}_\mathbf{j}(\mathbf{u} \setminus \mathbf{v_j})|^2} \end{aligned}$$

(with $P(\mathbf{S} = \mathbf{s}|\mathcal{A}_\mathbf{j}) = 0$ otherwise) From this, it follows (Jacobi, 1841) that $E_\lambda[\widehat{\boldsymbol{\beta}}_\mathbf{S}|\mathcal{A}_\mathbf{j}]$ is simply the least squares estimate of $\boldsymbol{\beta}$ based on the predictors whose indices lie in the set $\mathbf{j}$ and so

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\lambda) &= E_\lambda[\widehat{\boldsymbol{\beta}}_\mathbf{S}] \\ &= \sum_\mathbf{j} P(\mathcal{A}_\mathbf{j}) E_\lambda[\widehat{\boldsymbol{\beta}}_\mathbf{S}|\mathcal{A}_\mathbf{j}] \\ &= \sum_\mathbf{j} a_\lambda(\mathbf{j}) \widehat{\boldsymbol{\beta}}_\mathbf{j}. \end{aligned}$$

The form of $a_\lambda(\mathbf{j})$ now follows from Proposition 1. If $\mathbf{v_j}$ is non-empty we have

$$\begin{aligned} a_\lambda(\mathbf{j}) &= P(\mathcal{A}_\mathbf{j}) \\ &= P(\mathbf{S} \cap \{n+1, \cdots, n+p\} = \mathbf{v_j}) \\ &= |\mathcal{X}^T\mathcal{X} + \lambda I|^{-1} \lambda^{p-k} \sum_\mathbf{s} |\mathcal{X}_\mathbf{j}(\mathbf{s} \setminus \mathbf{v})|^2 \\ &= \frac{\lambda^{p-k}|\mathcal{X}(\mathbf{j})^T\mathcal{X}(\mathbf{j})|}{|\mathcal{X}^T\mathcal{X} + \lambda I|} \end{aligned}$$

where $\text{card}(\mathbf{j}) = k$. If $\mathbf{j} = \emptyset$ then $a_\lambda(\mathbf{j}) = \lambda^p|\mathcal{X}^T\mathcal{X} + \lambda I|^{-1}$ from Proposition 1. Note that for $\mathbf{j} = \{1, \cdots, p\}$, we have

$$\begin{aligned} a_\lambda(\mathbf{j}) &= \frac{|\mathcal{X}^T\mathcal{X}|}{|\mathcal{X}^T\mathcal{X} + \lambda I|} \\ &= \left|\mathcal{X}^T\mathcal{X}(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\right| \\ &= |I - \lambda(\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}|. \end{aligned}$$

References

Ali, A., Kolter, J.Z., Tibshirani, R.J.: A continuous-time view of early stopping for least squares. arXiv: 1810.10082 (2018)

Allen-Zhu, Z., Li, Y.: Faster principal component regression and stable matrix Chebyshev approximation. In: Proceedings of the 34th International Conference on Machine Learning. **70**, 107–115 (2017)

Artemiou, A., Li, B.: On principal components and regression: a statistical explanation of a natural phenomenon. Statistica Sinica. **19**, 1557–1565 (2009)

Attouch, H., Wets, R.J.B.: Approximation and convergence in nonlinear optimization. Nonlinear Programming. **4** (1981)

Berman, M.: A theorem of Jacobi and its generalization. Biometrika, **75**, 779–783 (1988)

Bernstein, S.: Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. Communications of the Kharkov Mathematical Society. **13**, 1–2 (1912)

Breiman, L.: Bagging predictors. Machine learning. **24**. 123–40 (1996)

Breiman, L.: Random forests. Machine learning. **45**, 5–32 (2001)

Bühlmann, P., Yu, B.: Boosting with the L_2 loss: regression and classification. Journal of the American Statistical Association, **98**, 324–339 (2003)

Choi, M.-D., Wu, P.Y.: Convex combinations of projections. Linear algebra and its applications. **136**, 25–42 (1990)

Cohen, A.: All admissible linear estimates of the mean vector. Annals of Mathematical Statistics, **37**, 458–463 (1966)

Cook, R.D.: Dimension reduction in regression. Statistical Science. **22**, 1–26 (2007)

Cox, D.R.: Notes on some aspects of regression analysis. Journal of the Royal Statistical Society, Series A. **131**, 265–279 (1968)

Efron, B.: How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association. **81**, 461–470 (1986)

Farnham, S.D., Shen, L., Suter, B.: Principal component projection with low-degree polynomials. arXiv: 1902.08656 (2019)

Freund, Y., Shapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: European Conference on Computational Learning, 23–37 (1995)

Frostig, R., Musco, C., Musco, C. Sidford, A.: Principal component projection with principal component analysis. In: Proceedings of the 33rd International Conference on Machine Learning. **48**, 2349–2357 (2016)

George, E., McCulloch, R.: Variable selection via Gibbs sampling. Journal of the American Statistical Association. **88**, 881–889 (1993)

Hastie, T., Montanari, A., Rosset, S., Tibshirani, R.: Surprises in high-dimensional ridgeless least squares interpolation. arXiv: 1903.08560 (2019)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer. (2001)

Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, **12**, 55-67 (1970)

Hoerl, R.W.: Ridge regression: a historical context. Technometrics, **62**, 420–425 (2020)

Hutchinson, M.: A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Communications in Statistics - Simulation and Computation. **19**, 433–450 (1990)

Jacobi, C.G.J.: De formatione et proprietatibus Determinatium. Journal für die reine und angewandte Mathematik, **22**, 285–318 (1841)

Janson, L., Fithian, W,, Hastie, T.: Effective degrees of freedom: a flawed metaphor. Biometrika, **102**, 479–485 (2015)

Kallenberg, O.: Random Measures. Akademie-Verlag. (1983)

Knight, K.: Elemental estimates, influence, and algorithmic leveraging. In: Nonparametric Statistics, 3rd ISNPS Avignon. 219–231 (2019)

Kobak, D., Lomond, J., Sanchez, B.: The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to implicit ridge regularization. Journal of Machine Learning Research, **21**, 1–16 (2020)

Leamer, E., Chamberlain, G.: A Bayesian interpretation of pretesting. Journal of the Royal Statistical Society, Series B. **38**, 85–94 (1976)

LeJeune, D., Javadi, H., Baraniuk, R.G.: The implicit regularization of ordinary least squares ensembles. arXiv: 1910.04743 (2019)

Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. **12**, 591–612 (1970)

Mayo, M.S., Gray, J.B.: Elemental subsets: the building blocks of regression. The American Statistician. **51**, 122–129 (1997)

Milanfar, P.: Symmetrizing smoothing filters. SIAM Journal of Imaging Sciences. **6**, 263–284 (2013)

Quenouille, M.: Approximate tests of correlation in time-series. Journal of the Royal Statistical Society, Series B. **11**, 68–84 (1949)

Quenouille, M.: Notes on bias in estimation. Biometrika. **43**, 353–360 (1956)

Raftery, A, Madigan, D., Hoeting, J.: Bayesian model averaging for linear regression. Journal of the American Statistical Association. **92**, 179–191 (1997)

Resnick, S.: Extreme values, regular variation and point processes. Springer. (2013)

Samson, S., Zatorre, R., Ramsay, J.: Multidimensional scaling of synthetic musical timbre: perception of spectral and temporal characteristics. Canadian Journal of Experimental Psychology. **51**, 307–315 (1997)

Schucany, W.R., Gray, H.L., Owen, D.B.: On bias reduction in estimation. Journal of the American Statistical Association. **66**, 524–533 (1971)

Skilling, J.: The eigenvalues of mega-dimensional matrices. In: Maximum Entropy and Bayesian Methods. 455–466 (1989)

Skouras, K., Goutis, C., Bramson, M.: Estimation in linear models using gradient descent with early stopping. Statistics and Computing. **4**, 271–278 (1994)

Stein, C.: Estimation of the mean of a multivariate normal distribution. Annals of Statistics. **9**, 1135–1151.

Subrahmanyam, M.: A property of simple least squares estimates. Sankhya, Series B. **34**, 355–356 (1972)

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. **58**, 267–288 (1996)

Tibshirani, R.J.: Degrees of freedom and model search. Statistica Sinica. **25**, 1265–1296 (2015)

Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley: Boston (1977)

Wolpert, D.H.: Stacked generalization. Neural networks **5**, 241–259 (1992)