- **Note:** To appear in *Recent Developments in Multivariate and Random Matrix Analysis: Festschrift in Honour of Dietrich von Rosen.* Thomas Holgersson and Martin Singull, Editors. Springer, pp. 245–254. DOI. Book's website.

On shrinkage estimators and "effective degrees of freedom"

Lynn R. LaMotte[*] Julia Volaufova[†] Simo Puntanen[‡]

Abstract

Explicit expressions for the estimated mean $\tilde{\boldsymbol{y}}_k = X\tilde{\boldsymbol{\beta}}_k = H_k\boldsymbol{y}$ and *effective degrees of freedom* $\nu_k = \mathrm{tr}(H_k)$ by penalized least squares, with penalty $k||D\boldsymbol{\beta}||^2$, can be found readily when $X'X + D'D$ is nonsingular. We establish them here in general under only the condition that X be a non-zero matrix, and we show that the monotonicity properties that are known when $X'X$ is nonsingular also hold in general, but that they are affected by estimability of $D\boldsymbol{\beta}$. We establish the relation between these penalized least squares estimators and least squares under the restriction that $D\boldsymbol{\beta} = \mathbf{0}$.

Keywords Penalized least squares, smoothing parameter, shrinkage estimator

AMS Classification 62F10, 62J07.

[*]Biostatistics Program, LSU Health-NO, School of Public Health, 2020 Gravier St., New Orleans, LA 70112 USA (`llamot@lsuhsc.edu`)

[†](`jvolau@lsuhsc.edu`)

[‡]Faculty of Information Technology and Communication Sciences, FI-33014 Tampere University, (`simo.puntanen@tuni.fi`)

1 Introduction

For an n-vector response $\boldsymbol{y}$ with mean vector $\boldsymbol{\mu}$, variance-covariance matrix $\sigma^2 \mathrm{I}_n$, and a model $X\boldsymbol{\beta}$, the least-squares (OLS) estimate of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-}X'\boldsymbol{y} = \mathbf{P}_X\boldsymbol{y} = H_0\boldsymbol{y}. \tag{1}$$

X is a fixed, known $n \times p$ matrix, $\boldsymbol{\beta}$ is a p-vector of unknown parameters, and $\hat{\boldsymbol{\beta}}$ satisfies $X\hat{\boldsymbol{\beta}} = H_0\boldsymbol{y}$. The dimension of the column space of X is $\nu_0 = \mathrm{rank}(X) = \mathrm{tr}(H_0)$: that is, the model provides ν_0 *degrees of freedom* to fit $\boldsymbol{\mu}$ to $\boldsymbol{y}$. To state the link directly, the dimension of the model space $\mathbf{R}(X) = \mathbf{R}(H_0)$ is its degrees of freedom.

The term *degrees of freedom* (used here as a singular noun) has been in use since the beginnings of applied statistics. It has also been used in linear algebra for the dimension of the solution set for a system of linear equations. Thus $SS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ has $n-1$ degrees of freedom because the n terms squared, $y_i - \bar{y}$, which satisfy the single equation $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$, can be represented linearly in terms of $n-1$ independent linear functions of $y_1, \ldots, y_n$.

Degrees of freedom shows up wherever a chi-squared random variable is involved, where it is the number of independent normal random variables squared and summed. Consequently it shows up in Student's t and F statistics. In all these uses, it can be directly defined as the dimension of a linear subspace, and hence also as the trace of an orthogonal projection matrix, that is, of a symmetric idempotent matrix.

Given a matrix D and constant $k \geq 0$, penalties for departures of $D\boldsymbol{\beta}$ from $\mathbf{0}$ can be introduced by regressing $\begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix}$ on $\begin{pmatrix} X \\ \sqrt{k}D \end{pmatrix}$ (see Allen (1974)). The penalized least squares estimate of $\boldsymbol{\mu}$ is then

$$\tilde{\boldsymbol{y}}_k = X\tilde{\boldsymbol{\beta}}_k = X(X'X + kD'D)^{-}X'\boldsymbol{y} = H_k\boldsymbol{y}. \tag{2}$$

Increasing the *smoothing parameter* k increases the penalty $k\boldsymbol{\beta}'D'D\boldsymbol{\beta}$ and so pushes $D\tilde{\boldsymbol{\beta}}_k$ toward zero. Other possibilities include replacing $\sqrt{k}D$ with $\sqrt{K}D$, with diagonal matrix $\sqrt{K} = \text{Diag}(\sqrt{k_i} \geq 0)$, permitting separate controls in multiple dimensions.

Often $\nu_k = \text{tr}(H_k)$ is called the *effective degrees of freedom* of the penalized model: see Hastie et al. (2009), p.153, for example. It is supposed to index, somehow, the complexity and flexibility of the model. Ruppert et al. (2003) say that it "has the rough interpretation as the *equivalent number of parameters* and can be calibrated with polynomial fits." These authors, and others, have noted that ν_k

Estimates like (2) have been called *shrinkage* estimates, because $||H_k\boldsymbol{y}||^2 \leq ||H_0\boldsymbol{y}||^2$. The notion of shrinkage dates back at least to Stein's seminal paper, Stein (1956). A later version, *ridge regression*, received widespread attention (see Alldredge and Gilb (1976)). It is (2) with $D = \text{I}$. In that case, $X'X + k\text{I}$ is nonsingular when $k > 0$, and its shrinkage property is apparent. It can be shown that ridge regression shrinkage is monotone: for $k_2 > k_1$, both $H_{k_1} - H_{k_2}$ and $H_{k_1}^2 - H_{k_2}^2$ are nnd, and hence $\nu_{k_2} < \nu_{k_1}$ and $||\tilde{\boldsymbol{y}}_{k_2}||^2 < ||\tilde{\boldsymbol{y}}_{k_1}||^2$. In the limit as $k \to \infty$, $\tilde{\boldsymbol{y}}_k \to \mathbf{0}$.

Extant accounts (Hastie et al. (2009), Ruppert et al. (2003), for example) implicitly assume that $X'X + kD'D$ is nonsingular in order to establish these properties of H_k, H_k^2, and ν_k, and their limits. Models widely used in practice entail X and D matrices such that $X'X + D'D$ is singular. Models that include effects of one or more treatment factors often are formulated with dummy, or indicator, variables, in which case the columns of the X matrix are linearly dependent. Further, main effects and interaction effects are formulated correspondingly, and imposing conditions like "no AB interaction effects" takes the form $D\boldsymbol{\beta} = \mathbf{0}$ with non-full-column-rank D. Similarly, differences (first order, second order) can also take the form $D\boldsymbol{\beta}$ with linearly dependent columns of D. Questions of estimability and effects arise when X has less than full column rank. For example, increasing penalties on $D\boldsymbol{\beta}$ in some directions may have no effect at all, as if D or k were zero.

Our objective here is to assess these same properties in general, when $X'X + D'D$ is not necessarily positive definite. We will show that ν_k and H_k are monotone decreasing functions of k, and we will show that $\tilde{\boldsymbol{y}}_k$ shrinks toward the least-squares estimate under the restriction that $D\tilde{\boldsymbol{\beta}} = \mathbf{0}$. Furthermore, we will show that the dimension of the space over which $\tilde{\boldsymbol{y}}_k$ ranges is the same for all k, and hence that ν_k has no relation to the dimension of the model space.

2 Propositions

Assume throughout that matrices (all real here) are conformable for the operations and expressions in which they appear. For matrices A and B, matrix sum, product, transpose, trace, generalized inverse, Moore-Penrose pseudoinverse, and inverse are denoted by $A+B$, AB, A', $\mathrm{tr}(A)$, A^-, A^+, and A^{-1}, respectively. For a symmetric, non-negative definite (nnd) matrix A, $A^{1/2}$ denotes a symmetric matrix such that $A^{1/2}A^{1/2} = A$. $\mathrm{Diag}(\alpha_i)$ denotes an $r \times r$ diagonal matrix with diagonal entries α_i, $i = 1, \ldots, r$. The orthogonal complement of a set $\mathcal{S}$ of vectors in $\Re^n$ is denoted $\mathcal{S}^\perp$. Denote the column space of M by $\mathbf{R}(M) = \{M\boldsymbol{x} : \boldsymbol{x} \in \Re^c\}$ and the null space of M by $\mathbf{N}(M) = \{\boldsymbol{x} \in \Re^c : M\boldsymbol{x} = \mathbf{0}\} = \mathbf{R}(M')^\perp$. $\mathbf{P}_A$ denotes the orthogonal projection matrix onto the column space of A.

Let X and D be matrices, both with $p \geq 1$ columns. We shall use D without $\sqrt{k}$ to simplify notation in the following propositions. As the results hold for any D, they hold for $\sqrt{k}D$, too.

Proposition 1. *$H = X(X'X+D'D)^-X'$ is invariant to the choice of generalized inverse, and $\mathbf{R}(H) = \mathbf{R}(X)$.*

Proof. Clearly $\mathbf{R}(X') \subset \mathbf{R}(X', D') = \mathbf{R}(X'X + D'D)$, so there exists a matrix M such that $X' = (X'X + D'D)M$. Then, with $(X'X + D'D)^\dagger$ an arbitrary generalized inverse of $(X'X+D'D)$,

$$\begin{aligned} H &= X(X'X+D'D)^{-}X' \\ &= M'(X'X+D'D)(X'X+D'D)^{-}(X'X+D'D)M \\ &= M'(X'X+D'D)M \\ &= M'(X'X+D'D)(X'X+D'D)^{\dagger}(X'X+D'D)M \\ &= X(X'X+D'D)^{\dagger}X', \text{ hence the invariance.} \end{aligned}$$

It is clear that $\mathbf{R}(H) \subset \mathbf{R}(X)$. Because $H = M'(X'X+D'D)M$ is nnd, $H\boldsymbol{z} = \mathbf{0}$ $\Longrightarrow (X'X+D'D)M\boldsymbol{z} = X'\boldsymbol{z} = \mathbf{0}$, and so it follows that $\mathbf{R}(H)^{\perp} \subset \mathbf{R}(X)^{\perp}$, and hence $\mathbf{R}(X) \subset \mathbf{R}(H)$. ■

Proposition 8 in Section 4 is a re-statement of Theorem 6.2.3, p. 122, in Rao and Mitra (1971). Here, with both $A = X'X$ and $B = D'D$ nnd, the condition that $\mathbf{R}(N'A) \subset \mathbf{R}(N'AN)$ is satisfied. Then there exists a nonsingular matrix T such that $T'X'XT = \Delta_1 = \text{Diag}(\delta_{1i})$ and $T'D'DT = \Delta_2 = \text{Diag}(\delta_{2i})$. Note that the diagonal entries of both Δ_1 and Δ_2 are nonnegative. Let $\Delta = \Delta_1 + \Delta_2$. Note that $T\Delta^{+}T'$ is a generalized inverse of $T'^{-1}\Delta T^{-1} = X'X + D'D$. It follows that

$$\begin{aligned} H &= X(T'^{-1}\Delta T^{-1})^{-}X' \\ &= XT\Delta^{+}T'X' \\ &= XT(\Delta_1+\Delta_2)^{+}T'X'. \end{aligned}$$

Keep in mind that when we replace D by $\sqrt{k}D$, only Δ_2 is affected, and it is replaced by $k\Delta_2$.

Rearrange the columns of T as $T = (T_{++}, T_{+0}, T_{0+}, T_{00})$ to correspond to column numbers $J_{++} = \{j : \delta_{1j} > 0 \text{ and } \delta_{2j} > 0\}$, $J_{+0} = \{j : \delta_{1j} > 0 \text{ and } \delta_{2j} = 0\}$, $J_{0+} = \{j : \delta_{1j} = 0 \text{ and } \delta_{2j} > 0\}$, and $J_{00} = \{j : \delta_{1j} = 0 \text{ and } \delta_{2j} = 0\}$, respectively. With $X \neq 0$, not both J_{++} and J_{+0} are empty; if $D \neq 0$, not both J_{++} and J_{0+} are empty. All other configurations are possible. With these conventions,

$$
\begin{aligned}
T'X'XT &= \begin{pmatrix} T'_{++}X'XT_{++} & T'_{++}X'XT_{+0} & T'_{++}X'XT_{0+} & T'_{++}X'XT_{00} \\ T'_{+0}X'XT_{++} & T'_{+0}X'XT_{+0} & T'_{+0}X'XT_{0+} & T'_{+0}X'XT_{00} \\ T'_{0+}X'XT_{++} & T'_{0+}X'XT_{+0} & T'_{0+}X'XT_{0+} & T'_{0+}X'XT_{00} \\ T'_{00}X'XT_{++} & T'_{00}X'XT_{+0} & T'_{00}X'XT_{0+} & T'_{00}X'XT_{00} \end{pmatrix} \\
&= \begin{pmatrix} \Delta_{1++} & 0 & 0 & 0 \\ 0 & \Delta_{1+0} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \qquad (3)
\end{aligned}
$$

Similar notation will be applied to submatrices of Δ_2, that is,

$$
T'D'DT = \Delta_2 = \begin{pmatrix} \Delta_{2++} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \Delta_{20+} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \qquad (4)
$$

From (3) we get immediately that $XT_{0+} = 0$ and $XT_{00} = 0$. Analogously, $DT_{+0} = 0$ and $DT_{00} = 0$.

Proposition 2.

$$
\begin{aligned}
X(X'X + D'D)^{-}X' &= XT(\Delta_1 + \Delta_2)^{+}T'X' \\
&= XT \begin{pmatrix} (\Delta_{1++} + \Delta_{2++})^{-1} & 0 & 0 & 0 \\ 0 & \Delta_{1+0}^{-1} & 0 & 0 \\ 0 & 0 & \Delta_{20+}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} T'X' \\
&= XT_{++}(\Delta_{1++} + \Delta_{2++})^{-1}T'_{++}X' \\
&\qquad + XT_{+0}\Delta_{1+0}^{-1}T'_{+0}X'. \qquad (5)
\end{aligned}
$$

Proposition 3. $\mathbf{N}(D'D) = \mathbf{R}(T_{+0}, T_{00})$.

Proof. Note that $(T_{+0}, T_{00})'D'D(T_{+0}, T_{00}) = 0 \Longrightarrow D'D(T_{+0}, T_{00}) = 0 \Longrightarrow \mathbf{R}(T_{+0}, T_{00}) \subset \mathbf{N}(D'D)$.

Let $\boldsymbol{u} \in \mathbf{N}(D'D)$. Since T is nonsingular, $\boldsymbol{u} = T\boldsymbol{v}$, with $\boldsymbol{v} = T^{-1}\boldsymbol{u}$; and $D'D\boldsymbol{u} = D'DT\boldsymbol{v} = \mathbf{0} \Longrightarrow \boldsymbol{v}'T'D'DT\boldsymbol{v} = \boldsymbol{v}'\Delta_2\boldsymbol{v} = \boldsymbol{v}'_{++}\Delta_{2++}\boldsymbol{v}_{++} + \boldsymbol{v}'_{0+}\Delta_{20+}\boldsymbol{v}_{0+} = 0 \Longrightarrow \boldsymbol{v}_{++} = \mathbf{0}$ and $\boldsymbol{v}_{0+} = \mathbf{0}$. Therefore $\boldsymbol{u} = T\boldsymbol{v} = T(\mathbf{0}', \boldsymbol{v}'_{+0}, \mathbf{0}', \boldsymbol{v}'_{00})' = T_{+0}\boldsymbol{v}_{+0} + T_{00}\boldsymbol{v}_{00} \in \mathbf{R}(T_{+0}, T_{00})$. ■

The linear subspace $\{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \Re^p \text{ and } D\boldsymbol{\beta} = \mathbf{0}\}$ is the restricted model $X\boldsymbol{\beta}$ under the condition that $D\boldsymbol{\beta} = \mathbf{0}$. The following proposition establishes that it is the same as $\mathbf{R}(XT_{+0})$.

Proposition 4. $\{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \Re^p \textit{ and } D\boldsymbol{\beta} = \mathbf{0}\} = \mathbf{R}(XT_{+0})$.

Proof. If $\boldsymbol{\mu} = X\boldsymbol{\beta}_0$ and $D\boldsymbol{\beta}_0 = \mathbf{0}$ then $\boldsymbol{\beta}_0 \in \mathbf{N}(D'D)$, and so there exist $\boldsymbol{v}_{+0}$ and $\boldsymbol{v}_{00}$ such that $\boldsymbol{\beta}_0 = T_{+0}\boldsymbol{v}_{+0} + T_{00}\boldsymbol{v}_{00}$, and hence $X\boldsymbol{\beta}_0 = XT_{+0}\boldsymbol{v}_{+0}$ because $XT_{00} = 0$.

If $\boldsymbol{\mu} \in \mathbf{R}(XT_{+0})$ then $\exists\ \boldsymbol{v}_{+0}$ such that $\boldsymbol{\mu} = XT_{+0}\boldsymbol{v}_{+0}$; and $DT_{+0} = \mathbf{0}$, so $\boldsymbol{\mu} \in \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \Re^p \text{ and } D\boldsymbol{\beta} = \mathbf{0}\}$. ■

Now replace D by $\sqrt{k}D$, $k \geq 0$, so that

$$\begin{aligned} H_k &= XT_{++}(\Delta_{1++} + k\Delta_{2++})^{-1}T'_{++}X' \\ &\qquad + XT_{+0}\Delta_{1+0}^{-1}T'_{+0}X', \end{aligned} \tag{6}$$

and

$$\nu_k = \mathrm{tr}(H_k) = p_{+0} + \mathrm{tr}[(\Delta_{1++} + k\Delta_{2++})^{-1}\Delta_{1++}], \tag{7}$$

where p_{+0} is the column dimension of T_{+0}.

Proposition 5.

$$\begin{aligned} \lim_{k\to\infty} H_k &= XT_{+0}\Delta_{1+0}^{-1}T'_{+0}X' = \mathbf{P}_{XT_{+0}} \textit{ and} \\ \lim_{k\to\infty} \nu_k &= p_{+0}. \end{aligned}$$

Proposition 6.

$$\frac{d}{dk} H_k = -X T_{++} (\Delta_{1++} + k\Delta_{2++})^{-2} \Delta_{2++} T'_{++} X',$$

and

$$\frac{d\nu_k}{dk} = \frac{d}{dk} \mathrm{tr}(H_k) = -\mathrm{tr}[(\Delta_{1++} + k\Delta_{2++})^{-2} \Delta_{1++} \Delta_{2++}].$$

Proposition 7. *If $\Delta_{2++} \neq 0$, then for any $k_2 > k_1 \geq 0$, both $H_{k_1} - H_{k_2}$ and $H^2_{k_1} - H^2_{k_2}$ are non-zero and nnd.*

3 Discussion

Some parts of the partition of T can be absent, depending on X and D. If columns of X are linearly independent, then both J_{0+} and J_{00} are void, and so $T = (T_{++}, T_{+0})$. If $D'D$ is pd, then both J_{+0} and J_{00} are void, and $T = (T_{++}, T_{0+})$. In general, whether or not X has full column rank, if $\mathbf{R}(D') \subset \mathbf{R}(X')$, it can be seen that $\delta_{1j} = 0 \Longrightarrow \delta_{2j} = 0$, and so J_{0+} is empty, and $T = (T_{++}, T_{+0}, T_{00})$. Then positive entries in Δ_2 occur only with positive entries of Δ_1. On the other hand, if none of $D\boldsymbol{\beta}$ is estimable, so that $\mathbf{R}(D') \cap \mathbf{R}(X') = \{\mathbf{0}\}$, it can be shown that $H_k = \mathbf{P}_X$ and $\tilde{\boldsymbol{y}}_k = \hat{\boldsymbol{y}}$ for all $k \geq 0$; positive entries in Δ_2 occur only with 0 entries in Δ_1. More generally, it can be shown that H_k is affected only by the *estimable part* of D, the part of $\mathbf{R}(D')$ that is contained in $\mathbf{R}(X')$.

Proposition 1 establishes that, for any $k \geq 0$, $\mathbf{R}(H_k) = \mathbf{R}(X)$, and hence the set of possibilities for the shrinkage estimator $\tilde{\boldsymbol{y}}_k$ is the same as the set of possibilities for the ordinary least squares estimator. The dimension of $\mathbf{R}(H_k)$ is $\nu_0 = \mathrm{rank}(X)$ for all k. Only at $k = 0$, when $\nu_k = \nu_0 = \mathrm{rank}(X)$, is ν_k the dimension of any linear subspace that appears in this setting.

The partitioning of T provides a full-column-rank reparametrization of the model $X\boldsymbol{\beta}$ that provides restricted model and full model estimates and sums of squares for the restriction $D\boldsymbol{\beta} = \mathbf{0}$. This follows from the fact that $X(T_{++}, T_{+0})$ has full column rank, $\mathbf{R}(X) = \mathbf{R}[X(T_{++}, T_{+0})]$ is the orthogonal sum of $\mathbf{R}(XT_{++})$ and $\mathbf{R}(XT_{+0})$, and that the latter is the restricted model $\{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \Re^p \text{ and } D\boldsymbol{\beta} = \mathbf{0}\}$. It follows that $\boldsymbol{y}'\mathbf{P}_{XT_{++}}\boldsymbol{y}$ is the numerator sum of squares, with p_{++} degrees of freedom, for the F-statistic that tests the testable part of $\mathrm{H}_0 : D\boldsymbol{\beta} = \mathbf{0}$.

The expression (6) and Propositions 5 and 7 establish that $\tilde{\boldsymbol{y}}_k$ shrinks monotonically toward $\mathbf{P}_{XT_{+0}}\boldsymbol{y}$, which is the least-squares estimate of $X\boldsymbol{\beta}$ under the restriction $D\boldsymbol{\beta} = \mathbf{0}$, by Proposition 4.

Both propositions 6 and 7 establish that ν_k is a strictly decreasing function of k if $\Delta_{2++} \neq 0$. At $k = 0$ it is $\nu_0 = \text{rank}(X)$, and it decreases with k to approach its limit from above, which is p_{+0}. Proposition 7 establishes the shrinkage property, that the squared norm of the estimate $\tilde{\boldsymbol{y}}_k$ is also strictly decreasing in k if $\Delta_{2++} \neq 0$.

It has been suggested in some settings (e.g., Hastie et al. (2009), p.158) that the value of k be chosen so that ν_k is some chosen value, say p_*. The expression for the derivative of ν_k with respect to k in Proposition 6 can be used in Newton's method to solve the nonlinear equation $\nu_k = p_*$.

The limiting behavior of H_k may seem abrupt: for all $k \geq 0$, $\mathbf{R}(H_k) = \mathbf{R}(X)$ and its rank stays undiminished as ν_0, but it approaches $\mathbf{P}_{XT_{+0}}$, which has rank p_{+0}. However, this reflects a fundamental property of closed convex cones of nnd matrices: in the relative interior (as here with finite k) of such a cone, all matrices have the same rank, and the rank can decrease only at the relative boundary. (See Lemma 2 in LaMotte (1977)).) Here, H_k follows a path within the relative interior, and it approaches a matrix on the relative boundary that has lesser rank.

4 Construction of T

Proposition 8 is a slight re-statement of Theorem 6.2.3 in Rao and Mitra (1971), p.122). The proof we have included shows in detail how to construct T.

Proposition 8. *Let A and B be symmetric $n \times n$ matrices, B nnd, and let columns of N comprise an orthonormal basis of $\mathbf{R}(N) = \mathbf{R}(B)^{\perp} = \{\boldsymbol{x} \in \Re^n : B\boldsymbol{x} = \mathbf{0}\}$. Let r denote the rank of B.*

(a) If $\mathbf{R}(N'A) = \mathbf{R}(N'AN)$ then there exists a nonsingular matrix T such that $T'AT = \Delta_1 = \mathrm{Diag}(\delta_{1i})$ and $T'BT = \Delta_2 = \mathrm{Diag}(\delta_{2i})$.

(b) If there exists a nonsingular matrix T such that $T'AT$ and $T'BT$ are both diagonal, then $\mathbf{R}(N'A) = \mathbf{R}(N'AN)$.

Proof of (a)

By spectral decomposition of B, there exists a matrix L such that $L'BL = \mathrm{I}_r$. (With $B = P\Lambda P' = (P_+, P_0)\begin{pmatrix} \Lambda_+ & 0 \\ 0 & 0 \end{pmatrix} P' = P_+\Lambda_+P_+'$, with $P'P = PP' = \mathrm{I}$, let $L = P_+\Lambda_+^{-1/2}$. P_+ has r columns and P_0 has $n-r$ columns.) Let $N = P_0$.

Let $S = ((\mathrm{I} - N(N'AN)^- N'A)L, N)$. Then, with M such that $N'A = N'ANM$ because $\mathbf{R}(N'A) = \mathbf{R}(N'AN)$,

$$\begin{aligned} N'A(\mathrm{I} - N(N'AN)^- N'A) &= N'A - N'AN(N'AN)^- N'ANM) \\ &= N'A - N'ANM = 0. \end{aligned}$$

Then

$$S'AS = \begin{pmatrix} L'(\mathrm{I} - AN(N'AN)^{-\prime}N') \\ N' \end{pmatrix} A((\mathrm{I} - N(N'AN)^- N'A)L, N);$$

and

$$\begin{aligned}
&L'(\mathrm{I} - AN(N'AN)^{-\prime}N') \\
&\times A(\mathrm{I} - N(N'AN)^{-}N'A)L \;=\; L'AL \\
&\qquad -L'AN(N'AN)^{-\prime}N'AL - L'AN(N'AN)^{-}N'AL \\
&\qquad +L'AN(N'AN)^{-\prime}N'AN(N'AN)^{-}N'AL \\
&\;=\; L'AL - L'AN(N'AN)^{-}N'AL
\end{aligned}$$

because all the terms involving $AN(N'AN)^{-}N'A$ are invariant to the choice of generalized inverse, and $(N'AN)^{-\prime}$ is a generalized inverse of $N'AN$. Then

$$S'AS = \begin{pmatrix} L'(A - AN(N'AN)^{-}N'A)L & 0 \\ 0 & N'AN \end{pmatrix}.$$

Because $BN = 0$, $B(\mathrm{I} - N(N'AN)^{-}N'A) = B$, so that

$$L'(\mathrm{I} - AN(N'AN)^{-\prime}N')B((\mathrm{I} - N(N'AN)^{-}N'A)L = L'BL = \mathrm{I}_r.$$

Then

$$S'BS = \begin{pmatrix} L'BL & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathrm{I}_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Turning back to $S'AS$, the two matrices on the diagonal are symmetric, and so there exist orthonormal matrices M and Q such that

$$\begin{aligned}
M'L'(A - AN(N'AN)^{-}N'A)LM &= \Phi_1 \text{ and} \\
Q'N'ANQ &= \Phi_2,
\end{aligned}$$

where both Φ_1 and Φ_2 are diagonal. Let

$$\begin{aligned}
T &= S\begin{pmatrix} M & 0 \\ 0 & Q \end{pmatrix} \\
&= ((\mathrm{I} - N(N'AN)^{-}N'A)LM, NQ).
\end{aligned}$$

Then

$$\begin{aligned} T'AT &= \begin{pmatrix} M' & 0 \\ 0 & Q' \end{pmatrix} S'AS \begin{pmatrix} M & 0 \\ 0 & Q \end{pmatrix} \\ &= \begin{pmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{pmatrix} = \Delta_1, \end{aligned}$$

and

$$\begin{aligned} T'BT &= \begin{pmatrix} M' & 0 \\ 0 & Q' \end{pmatrix} S'BS \begin{pmatrix} M & 0 \\ 0 & Q \end{pmatrix} \\ &= \begin{pmatrix} M' & 0 \\ 0 & Q' \end{pmatrix} \begin{pmatrix} \mathrm{I}_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & Q \end{pmatrix} \\ &= \begin{pmatrix} M'M & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{I} & 0 \\ 0 & 0 \end{pmatrix} = \Delta_2. \end{aligned}$$

To show that the columns of T are linearly independent, suppose that $T\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \mathbf{0}$. Then

$$\begin{aligned} T\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} &= (\mathrm{I} - N(N'AN)^{-}N'A)LM\boldsymbol{u} + NQ\boldsymbol{v} \\ &= \mathbf{0} \end{aligned}$$

implies that

$$\begin{aligned} BT\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} &= B(\mathrm{I} - N(N'AN)^{-}N'A)LM\boldsymbol{u} + BNQ\boldsymbol{v} \\ &= BLM\boldsymbol{u} = \mathbf{0} \end{aligned}$$

because $BN = 0$. That implies that $\boldsymbol{u}'M'L'BLM\boldsymbol{u} = 0 \Longrightarrow \boldsymbol{u}'\boldsymbol{u} = 0$ because $M'L'BLM = \mathrm{I}$, hence $\boldsymbol{u} = \mathbf{0}$. That leaves $NQ\boldsymbol{v} = \mathbf{0}$, which implies that $\boldsymbol{v}'Q'N'NQ\boldsymbol{v} = \boldsymbol{v}'Q'Q\boldsymbol{v} = \boldsymbol{v}'\boldsymbol{v} = 0$, which implies that $\boldsymbol{v} = \mathbf{0}$. ■

For the proof of (b), let T be a nonsingular matrix such that $T'AT = \Delta_1$ and $T'BT = \Delta_2$, where both Δ_1 and Δ_2 are diagonal. Order columns of $T = (T_+, T_0)$ so that

$$\begin{aligned} T'BT &= \begin{pmatrix} T_+'BT_+ & T_+'BT_0 \\ T_0'BT_+ & T_0'BT_0 \end{pmatrix} \\ &= \begin{pmatrix} \Delta_{2+} & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} T'AT &= \begin{pmatrix} T_+'AT_+ & T_+'AT_0 \\ T_0'AT_+ & T_0'AT_0 \end{pmatrix} \\ &= \begin{pmatrix} \Delta_{1+} & 0 \\ 0 & \Delta_{10} \end{pmatrix}. \end{aligned}$$

Note that $\mathbf{R}(T_0) = \mathbf{R}(B)^\perp$

[Proof: That B is symmetric and nnd $\Longrightarrow T_0'BT_0 = \mathbf{0} \Longrightarrow BT_0 = 0$. Therefore $\mathbf{R}(T_0) \subset \mathbf{R}(B)^\perp$. With T nonsingular, $\boldsymbol{z} \in \mathbf{R}(B)^\perp \Longrightarrow \exists\, \boldsymbol{x}, \boldsymbol{w}$ such that $B\boldsymbol{z} = \mathbf{0} = B(T_0\boldsymbol{x} + T_+\boldsymbol{w}) = BT_+\boldsymbol{w} \Longrightarrow \boldsymbol{w}'T_+'BT_+\boldsymbol{w} = 0 \Longrightarrow \boldsymbol{w}'\Delta_{2+}\boldsymbol{w} = 0 \Longrightarrow \boldsymbol{w} = \mathbf{0}$ $\Longrightarrow \boldsymbol{z} = T_0\boldsymbol{x} \in \mathbf{R}(T_0)$.]

and $R(B) = R(BT_+)$. Define N to be $T_0(T_0'T_0)^{-1/2}$, so that $N'N = \mathrm{I}$ and $\mathbf{R}(N) = \mathbf{R}(B)^\perp$.

Proof of (b). With A, B, $T = (T_+, T_0)$, and N as defined above, proposition (b) is equivalent to

$$\mathbf{R}(N'AN)^\perp = \mathbf{R}(N'A)^\perp.$$

Clearly $\mathbf{R}(N'AN) \subset \mathbf{R}(N'A)$, and so it remains to prove that, for any $\boldsymbol{x} \in \Re^{n-r}$, $N'AN\boldsymbol{x} = 0 \Longrightarrow AN\boldsymbol{x} = \mathbf{0}$.

Suppose $N'AN\boldsymbol{x} = \mathbf{0}$. Then $AN\boldsymbol{x} \in \mathbf{R}(N)^\perp = \mathbf{R}(B) = \mathbf{R}(BT_+) \Longrightarrow \exists\, \boldsymbol{w}$ such that

$$AN\boldsymbol{x} = BT_+\boldsymbol{w}.$$

Then

$$\boldsymbol{w}'T_+'BT_+\boldsymbol{w} = \boldsymbol{w}'T_+'AN\boldsymbol{x} = 0,$$

because $T_+'AN = 0$; and this implies that $BT_+\boldsymbol{w} = \mathbf{0}$ because B is nnd. Therefore $AN\boldsymbol{x} = \mathbf{0}$. ■

Given A and B, defining T entails the following steps.

1. Spectrally decompose B to get P_+, Λ_+, $L = P_+\Lambda_+^{-1/2}$, $N = P_0$, and $(N'AN)^-$. Check whether $\mathbf{R}(N'A) \subset \mathbf{R}(N'AN)$ (one way is whether $(N'AN)(N'AN)^-N'A = N'A$).

2. Spectrally decompose $L'(A - AN(N'AN)^-N'A)L$ and $N'AN$ to get M and Q.

3. Compute

$$T = ((\mathrm{I} - N(N'AN)^-N'A)LM, NQ).$$

4. For $A = X'X$ and $B = D'D$, both nnd, compute $T'AT = \Delta_1$ and $T'BT = \Delta_2$; permute the columns of T to correspond to the sets J_{++}, J_{+0}, J_{0+}, and J_{00} that are non-void.

References

Alldredge J.R. and Gilb N.S. (1976). Ridge regression: an annotated bibliography, International Statistical Review 44(3), 355–360.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction, Technometrics 16 (1), 125–127.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Second edition. Springer Science+Business Media, LLC, New York.

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

LaMotte, L.R. (1977). On admissibility and completeness of linear unbiased estimators in a general linear model, J. Amer. Statist. Assoc. 72(358), 438–441.

Rao C.R. and Mitra S.K. (2015). Generalized inverse of matrices and its applications. John Wiley & Sons Inc.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric regression. Cambridge University Press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution,Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 197–206.