Electronic Journal of Statistics
Vol. 7 (2013) 1806–1834
ISSN: 1935-7524
DOI: 10.1214/13-EJS826

Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance

Lee H. Dicker*

Department of Statistics and Biostatistics
Rutgers University
501 Hill Center, 110 Frelinghuysen Road
Piscataway, NJ 08854
e-mail: ldicker@stat.rutgers.edu

Abstract: In a linear model, consider the class of estimators that are equivariant with respect to linear transformations of the predictor basis. Each of these estimators determines an equivariant linear prediction rule. Equivariant prediction rules may be appropriate in settings where sparsity assumptions (like those common in high-dimensional data analysis) are untenable or little is known about the relevance of the given predictor basis, insofar as it relates to the outcome. In this paper, we study the out-of-sample prediction error associated with equivariant estimators in high-dimensional linear models with Gaussian predictors and errors. We show that non-trivial equivariant prediction is impossible when the number of predictors d is greater than the number of observations n. For $d/n \to \rho \in [0, 1)$, we show that a James-Stein estimator (a scalar multiple of the ordinary least squares estimator) is asymptotically optimal for equivariant out-of-sample prediction, and derive a closed-form expression for its asymptotic predictive risk. Finally, we undertake a detailed comparative analysis involving the proposed James-Stein estimator and other well-known estimators for non-sparse settings, including the ordinary least squares estimator, ridge regression, and other James-Stein estimators for the linear model. Among other things, this comparative analysis sheds light on the role of the population-level predictor covariance matrix and reveals that other previously studied James-Stein estimators for the linear model are sub-optimal in terms of out-of-sample prediction error.

AMS 2000 subject classifications: Primary 62J07; secondary 62C99, 62J05.
Keywords and phrases: Adaptive estimation, James-Stein estimator, out-of-sample prediction, oracle estimator, ridge regression.

Received January 2013.

1. Introduction

Consider a linear model with observed outcomes $y_1, \ldots, y_n \in \mathbb{R}$ and corresponding d-dimensional predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. The outcomes and predictors are related via

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

*Supported by NSF Grant DMS-1208785.

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ is an unknown parameter and $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}$ are unobserved errors. To simplify notation, let $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$. Then (1) may be rewritten as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

High-dimensional linear models, where d is large, have been extensively studied in recent research. In this challenging setting, additional conditions on $\boldsymbol{\beta}$, such as sparsity, are often required in order to ensure consistent estimation or prediction. Two of the more widely studied types of sparsity are ℓ^0- and ℓ^1-sparsity: $\boldsymbol{\beta}$ is sparse if its ℓ^0- or ℓ^1-norm is "small." While sparsity conditions may be required for consistency in high-dimensional linear models, these conditions may be untenable in some instances. Moreover, it remains important to identify optimal methods for practical objectives like out-of-sample prediction, even in non-sparse (or "dense") high-dimensional settings.

In this paper, we study high-dimensional out-of-sample prediction problems and a class of estimators that are equivariant with respect to linear transformations of the predictors $\mathbf{x}_i$, under the assumption that the data are multivariate normal. We argue that equivariant estimators are appropriate in problems where there is little prior knowledge about the relevance of the given predictor basis vis-à-vis the outcome y_i. In particular, equivariant estimators may be appropriate in settings where sparsity assumptions on $\boldsymbol{\beta}$ are not desirable or realistic, as sparsity is highly dependent on the predictor basis. Our analysis provides new insight into the capabilities and limitations of dense methods for high-dimensional linear models.

Most of the results in this paper fall into one of three categories: (i) impossibility results, (ii) optimality results, and (iii) comparative results. Our primary impossibility result (Theorem 1 (c) in Section 3) implies that if $d > n$, then the *only* equivariant estimator is $\hat{\boldsymbol{\beta}}_{null} = 0$; thus, non-trivial equivariant prediction is impossible when $d > n$. While it is widely understood that high-dimensional "dense problems" are very difficult, our impossibility results help to make this idea more precise. It is worth pointing out that these results are derived under the assumption that $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is an arbitrary, unknown positive definite matrix (this is a random predictor analysis). If Σ is known or estimable, then results from (Dicker, 2013) imply that non-trivial equivariant prediction may be possible even when $d > n$; this is discussed in more detail in Section 5.2.

The optimality results in this paper primarily focus on a class of James-Stein estimators for $\boldsymbol{\beta}$, which are scalar multiples of the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{ols} = (X^TX)^{-1}X^T\mathbf{y}$. Stein shrinkage and the James-Stein estimator (James and Stein, 1961; Stein, 1955) are of fundamental importance in modern statistics. Most research on Stein shrinkage has focused on the Gaussian sequence model and the normal means problem; however, variants of the James-Stein estimator for linear models have also been studied (Baranchik, 1973; Copas, 1983; Huber and Leeb, 2012; Stein, 1960). The James-Stein estimators proposed in this paper are, to our knowledge, new. In Theorems 2-3 of Section 4, we prove that the proposed James-Stein estimators are asymptotically optimal among equivariant estimators when $d/n \to \rho \in [0, 1)$ and $d \to \infty$. Our

analysis of the James-Stein estimator shares similarities with (Marchand, 1993) and (Beran, 1996), who considered Stein estimation and equivariance in the normal means problem. However, the present analysis reveals unique features of linear models; for instance, our results demonstrate that adjusting for the degrees of freedom lost to a high-dimensional predictor has a non-trivial effect on out-of-sample prediction error.

Finally, we undertake a comparative analysis involving the proposed James-Stein estimator and other well-known dense estimators, including the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{ols} = (X^TX)^{-1}X^T\mathbf{y}$, ridge regression (Hoerl and Kennard, 1970; Tikhonov, 1943), and other previously studied James-Stein estimators for $\boldsymbol{\beta}$. In Theorem 4 of Section 5.1, we show that if $0 < \inf d/n \leq \sup d/n < 1$ and d, n are sufficiently large, then the proposed James-Stein estimator has uniformly smaller predictive risk than the OLS estimator; hence, under the specified conditions, the James-Stein estimator is minimax. Our discussion of ridge regression helps clarify the role of the predictor covariance matrix $\text{Cov}(\mathbf{x}_i) = \Sigma$ in dense out-of-sample prediction problems. In particular, we show that if Σ is known, then a certain equivariant ridge estimator has smaller predictive risk than the James-Stein estimator; furthermore, results from (Dicker, 2013) imply that this ridge estimator is asymptotically optimal among equivariant estimators *that may depend on* Σ. More standard ridge estimators (which do not require knowledge of Σ) are also discussed. After discussing ridge regression, we consider another previously studied James-Stein estimator for $\boldsymbol{\beta}$ (Baranchik, 1973) and show that – perhaps surprisingly – it is sub-optimal, in terms of out-of-sample prediction error.

The rest of the paper is organized as follows. In Section 2, we introduce notation and definitions, and describe the statistical setting for what follows. Equivariance is discussed in Section 3. The James-Stein estimator is defined in Section 4; some of its optimality properties are also discussed in Section 4. Section 5 contains a comparative analysis of the proposed James-Stein estimator and other dense estimators for $\boldsymbol{\beta}$. A concluding discussion is contained in Section 6, where we consider practical implications of the results contained in this paper and possible extensions. Proofs may be found in the Appendices.

2. Notation, definitions, and the statistical setting

Let $PD(d)$ denote the collection of $d \times d$ positive definite matrices. In addition to assuming the linear model (1), we assume that

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} N(0, \Sigma) \text{ and } \epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2), \tag{2}$$

are independent, where $\Sigma \in PD(d)$ and $\sigma^2 > 0$. Linear models with similar distributional assumptions have been previously studied by Stein (1960), Baranchik (1973), Breiman and Freedman (1983), Brown (1990), and Leeb (2009), among others; Dicker (2013) considered the model (1)-(2) under the additional assumptions $\Sigma = I$ and $\sigma^2 = 1$. The significance of the normality assumption (2) – and the possibility of relaxing it – is further discussed in Section 6.1.

Each estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ determines a linear prediction rule, $\hat{y}(\mathbf{x}) = \mathbf{x}^T\hat{\boldsymbol{\beta}}$. The unconditional out-of-sample prediction error (predictive risk) of $\hat{\boldsymbol{\beta}}$ is given by

$$E\{y_{new} - \hat{y}(\mathbf{x}_{new})\}^2 = E(y_{new} - \mathbf{x}_{new}^T\hat{\boldsymbol{\beta}})^2 = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sigma^2, \quad (3)$$

where $(y_{new}, \mathbf{x}_{new}^T)$ is independent of $(\mathbf{y}, X)$ and has the same distribution as $(y_i, \mathbf{x}_i^T)$. We emphasize that the expectation in (3) is taken over $(y_{new}, \mathbf{x}_{new})$ and $(\mathbf{y}, X)$. Broadly speaking, the goal of the unconditional out-of-sample prediction problem considered in this paper is to minimize (3) over estimators $\hat{\boldsymbol{\beta}}$.

In order to introduce more convenient notation for studying out-of-sample prediction error, let $\mathbf{w}_i = (y_i, \mathbf{x}_i^T) \in \mathbb{R}^{d+1}$. Then the assumption (2) is equivalent to assuming $\mathbf{w}_1, \ldots, \mathbf{w}_n \overset{\text{iid}}{\sim} N(0, V)$, where

$$V = \begin{pmatrix} \sigma^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} & \boldsymbol{\beta}^T\Sigma \\ \Sigma\boldsymbol{\beta} & \Sigma \end{pmatrix} \in PD(d+1). \quad (4)$$

In this way, we establish a correspondence between the parameters $\Sigma \in PD(d)$, $\boldsymbol{\beta} \in \mathbb{R}^d$, and $\sigma^2 > 0$ in the linear model (1), and positive definite matrices $V \in PD(d+1)$. After standardizing by σ^2, the predictive risk (3) is equivalent to

$$R_V(\hat{\boldsymbol{\beta}}) = \sigma^{-2}E_V\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\},$$

where the subscript V in the expectation on the right-hand side above indicates that the expectation is taken over $\mathbf{w}_1, \ldots, \mathbf{w}_n \sim N(0, V)$. In fact, $R_V(\hat{\boldsymbol{\beta}})$ is the primary object of study in the sequel and we will typically refer to $R_V(\hat{\boldsymbol{\beta}})$ itself as the predictive risk (or out-of-sample prediction error) of $\hat{\boldsymbol{\beta}}$. Note that the predictive risk $R_V(\hat{\boldsymbol{\beta}})$ is completely determined by the estimator $\hat{\boldsymbol{\beta}}$ and the positive definite matrix $V \in PD(d+1)$. We will often write $E_\Sigma(\cdot)$ in place of $E_V(\cdot)$ when the expectation only involves the random predictors X. Similarly, we write $P_V(\cdot)$ or $P_\Sigma(\cdot)$ when computing probabilities involving $\mathbf{w}_1, \ldots, \mathbf{w}_n$ or X, respectively.

3. Equivariance

Consider the following definition.

Definition 1. A measurable estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X)$ is *linearly equivariant* if

$$A\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}(\mathbf{y}, XA^{-1}) \quad (5)$$

for all $d \times d$ invertible matrices $A \in GL(d)$. It is *scale invariant* if

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}(t\mathbf{y}, tX) \quad (6)$$

for all positive scalars $t > 0$. If an estimator is both linearly equivariant and scale invariant, we say that it is *LiSc*.

As an initial example, notice that the OLS estimator is LiSc. Linearly equivariant estimators are compatible with linear transformations of the predictors $\mathbf{x}_i$. Intuitively, this type of compatibility implies that the data are treated "the same" (for the purposes of prediction), regardless of the given predictor basis. Hence, linearly equivariant estimators may be preferred in situations where there is little prior knowledge about the relevance of the given predictor basis, insofar as it relates to the outcome. By contrast, sparsity assumptions convey specific information about the designated predictor basis, and linear equivariance is less appropriate for sparse problems. Indeed, most sparse estimators, such as lasso (Tibshirani, 1996), are *not* linearly equivariant. Scale invariance is less specific to non-sparse problems; however, in our view, it is a reasonable property to require of estimators for $\boldsymbol{\beta}$, including sparse estimators (see, for instance, the scaled lasso (Sun and Zhang, 2012)).

In this paper, we primarily focus on LiSc estimators. Our main objectives include (i) finding LiSc estimators with small predictive risk and (ii) understanding the magnitude of these estimators' predictive risk in high-dimensional linear models.

A nice feature of LiSc estimators is that their predictive risk is completely determined by the signal-to-noise ratio (in addition to d, n). In particular, in order to evaluate the predictive risk of an LiSc estimator, it suffices to consider the case where $\Sigma = I$; this greatly simplifies calculations involving LiSc estimators. Define

$$\Theta^d(\eta^2) = \{V \in PD(d+1);\ \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}/\sigma^2 = \eta^2\}, \quad \eta \geq 0,$$

where the relationship between V and $\boldsymbol{\beta}$, Σ, σ^2 is given by (4). Then $\Theta^d(\eta^2)$ is the class of linear models with signal-to-noise ratio $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}/\sigma^2 = \eta^2$. The following proposition is proved in Appendix A.

Proposition 1. *Suppose that $V \in \Theta^d(\eta^2)$.*

(a) If $\hat{\boldsymbol{\beta}}$ is linearly equivariant, then $R_V(\hat{\boldsymbol{\beta}}) = R_{V_0}(\hat{\boldsymbol{\beta}})$, where

$$V_0 = \begin{pmatrix} \sigma^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} & \boldsymbol{\beta}^T\Sigma^{1/2} \\ \Sigma^{1/2}\boldsymbol{\beta} & I \end{pmatrix} \in \Theta^d(\eta^2).$$

(b) If $\hat{\boldsymbol{\beta}}$ is LiSc, then $R_V(\hat{\boldsymbol{\beta}}) = R_{V_{\mathbf{u}}}(\hat{\boldsymbol{\beta}})$, where

$$V_{\mathbf{u}} = \begin{pmatrix} 1+\eta^2 & \eta\mathbf{u}^T \\ \eta\mathbf{u} & I \end{pmatrix} \in \Theta^d(\eta^2)$$

and $\mathbf{u} \in \mathbb{R}^d$ is any fixed unit vector.

Proposition 1 indicates that the signal-to-noise ratio $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}/\sigma^2$ plays an important role in the analysis of LiSc estimators. The signal-to-noise ratio's significance is further highlighted in our subsequent analysis of LiSc estimators, where we first focus on "oracle" estimators, which are derived under the assumption that the signal-to-noise ratio is known, and then study "adaptive" estimators, which rely on an estimate of the signal-to-noise ratio.

Define the collection of LiSc estimators,

$$\mathscr{E} = \{\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X) \text{ is a LiSc estimator for } \boldsymbol{\beta}\},$$

and the optimal LiSc risk for $V \in \Theta^d(\eta^2)$,

$$r(\eta^2) = \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} R_V(\hat{\boldsymbol{\beta}}).$$

By Proposition 1, $r(\eta^2)$ is well-defined.

LiSc estimators have a great deal of structure. By taking advantage of this structure, we are able to identify optimal LiSc estimators in settings where $d < n$, $d = n$, and $d > n$, separately, assuming that $\eta^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2$ is known.

Theorem 1. *Let $V \in \Theta^d(\eta^2)$.*

(a) Suppose that $d < n$. Define

$$\hat{\eta}^2 = \frac{||\mathbf{y}||^2}{n\hat{\sigma}^2} - 1 \quad \text{and} \quad h_{opt}(\hat{\eta}^2) = \frac{E_V(\boldsymbol{\beta}^T \Sigma \hat{\boldsymbol{\beta}}_{ols} | \hat{\eta}^2)}{E_V(\hat{\boldsymbol{\beta}}_{ols}^T \Sigma \hat{\boldsymbol{\beta}}_{ols} | \hat{\eta}^2)}, \tag{7}$$

where $\hat{\sigma}^2 = (n-d)^{-1}||\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ols}||^2$. Then $h_{opt}(\hat{\eta}^2)$ is completely determined by η^2 and $\hat{\eta}^2$ (along with d, n), and

$$R_V(\hat{\boldsymbol{\beta}}_{opt}) = r(\eta^2),$$

where $\hat{\boldsymbol{\beta}}_{opt} = h_{opt}(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols}$.

(b) Suppose that $d = n$. If $\hat{\boldsymbol{\beta}}$ is LiSc, then there exists a constant $c \in \mathbb{R}$ and an LiSc estimator $\hat{\boldsymbol{\beta}}_c \triangleq cX^{-1}\mathbf{y}$ such that $R_V(\hat{\boldsymbol{\beta}}_c) \leq R_V(\hat{\boldsymbol{\beta}})$. Furthermore,

$$R_V(\hat{\boldsymbol{\beta}}_c) = \begin{cases} \infty & \text{if } c \neq 0 \\ \eta^2 & \text{if } c = 0 \end{cases}$$

and $r(\eta^2) = R_V(\hat{\boldsymbol{\beta}}_{null}) = \eta^2$, where $\hat{\boldsymbol{\beta}}_{null} = \hat{\boldsymbol{\beta}}_0 = 0$.

(c) Suppose that $d > n$. If $\hat{\boldsymbol{\beta}}$ is LiSc, then $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{null} = 0$.

Theorem 1 is proved in Appendix A. Theorem 1 (b)-(c) implies that when $d \geq n$, the optimal LiSc estimator is $\hat{\boldsymbol{\beta}}_{null} = 0$. Theorem 1 (c) implies further that if $d > n$, then $\hat{\boldsymbol{\beta}}_{null}$ is the *only* LiSc estimator. Thus, we have a fairly definitive characterization of LiSc estimators for out-of-sample prediction when $d \geq n$. This characterization is quite negative, which may raise questions about the appropriateness of LiSc estimators for high-dimensional data analysis. We defer such questions to Section 6.2, which contains a broader discussion of LiSc estimators and high-dimensional data analysis.

If $d < n$, then the optimal LiSc estimator is nontrivial and is given by $\hat{\boldsymbol{\beta}}_{opt} = h_{opt}(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols}$ in Theorem 1 (a). Theorem 1 (a) should be compared with Section 2.1 of (Marchand, 1993), where a best equivariant estimator for the normal means problem is derived. Observe that evaluating $h_{opt}(\hat{\eta}^2)$ seems

challenging and $h_{opt}(\hat{\eta}^2)$ depends on the signal-to-noise ratio, which is typically unknown; thus, implementing $\hat{\boldsymbol{\beta}}_{opt}$ in practice is generally infeasible. Furthermore, Theorem 1 (a) does not provide any information about the magnitude of $r(\eta^2)$, which is important for understanding the performance limits of LiSc estimators. All of these issues are addressed in the next section.

4. James-Stein estimators

In this section we study James-Stein shrinkage estimators for $\boldsymbol{\beta}$ and show that their predictive risk is asymptotically equivalent to the optimal LiSc risk $r(\eta^2)$ in high-dimensional linear models with $d/n \to \rho \in [0, 1)$ and $d \to \infty$. In Section 4.1, we identify an oracle James-Stein estimator that utilizes a non-random shrinkage parameter, which depends on the signal-to-noise ratio. We show that the oracle James-Stein estimator is asymptotically equivalent to the optimal LiSc estimator $\hat{\boldsymbol{\beta}}_{opt}$, which was derived in Theorem 1 (a). We also obtain an explicit formula for the predictive risk of the oracle James-Stein estimator; combined with our optimality results for the oracle James-Stein estimator, this easily yields an exact asymptotic expression for the minimal LiSc risk $r(\eta^2)$. In Section 4.2 we propose an adaptive James-Stein estimator that depends on a data-driven shrinkage parameter; this estimator more closely resembles the original James-Stein estimator (James and Stein, 1961), which also relies on a data-driven shrinkage parameter. We show that if $d/n \to \rho \in (0, 1)$, then the adaptive James-Stein estimator is asymptotically equivalent to the oracle estimator and, hence, the optimal LiSc estimator.

4.1. The oracle estimator

For $d < n-1$ and a shrinkage parameter $t \geq 0$, define the James-Stein estimator

$$\hat{\boldsymbol{\beta}}_{js}(t) = \frac{t}{t + d/(n-d-1)}(X^T X)^{-1} X^T \mathbf{y}.$$

Notice that for fixed $t \geq 0$, $\hat{\boldsymbol{\beta}}_{js}(t)$ is LiSc. Additionally, $\hat{\boldsymbol{\beta}}_{js}(0) = \hat{\boldsymbol{\beta}}_{null} = 0$ and $\hat{\boldsymbol{\beta}}_{js}(\infty) = \hat{\boldsymbol{\beta}}_{ols}$. A closed-form expression for the predictive risk of $\hat{\boldsymbol{\beta}}_{js}(t)$ follows easily from properties of the inverse-Wishart distribution; it is then straightforward to optimize over $t \geq 0$ and find the James-Stein estimator with minimal predictive risk. Details are given in the following proposition, which is proved in Appendix A.

Proposition 2. *Suppose that $d < n-1$ and that $V \in \Theta^d(\eta^2)$. If $t \geq 0$, then*

$$R_V\{\hat{\boldsymbol{\beta}}_{js}(t)\} = \left\{\frac{t}{t + d/(n-d-1)}\right\}^2 \frac{d}{n-d-1} + \left\{\frac{d/(n-d-1)}{t + d/(n-d-1)}\right\}^2 \eta^2 \quad (8)$$

and

$$R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} = \inf_{t \in [0,\infty]} R_V\{\hat{\boldsymbol{\beta}}_{js}(t)\} = \frac{\eta^2 d/(n-d-1)}{\eta^2 + d/(n-d-1)}.$$

Proposition 2 implies that the predictive risk of $\hat{\boldsymbol{\beta}}_{js}(t)$ is minimized when $t = \eta^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2$ is the signal-to-noise ratio. Since $\eta^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2$ is typically unknown, we refer to $\hat{\boldsymbol{\beta}}_{js}(\eta^2)$ as the "oracle James-Stein estimator." The next theorem relates the risk of the oracle James-Stein estimator to the minimal LiSc risk $r(\eta^2)$.

Theorem 2. *Suppose that* $\eta^2 \geq 0$ *and that* $0 < d/n \leq \rho_+ < 1$ *for some fixed constant* $\rho_+ \in \mathbb{R}$. *Then*

$$\sup_{V \in \Theta^d(\eta^2)} \left| R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} - r(\eta^2) \right| = O\left(\frac{\eta^2 d/n}{\eta^2 + d/n} d^{-1/2} \right). \tag{9}$$

Theorem 2 is proved in Appendix A. By Proposition 1, $|R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} - r(\eta^2)|$ is in fact constant over $V \in \Theta^d(\eta^2)$. Theorem 2 implies that if $n \to \infty$ and $\sup d/n < 1$, then the predictive risk of $\hat{\boldsymbol{\beta}}_{js}(\eta^2)$ is close to $r(\eta^2)$; in other words, the oracle James-Stein estimator is asymptotically optimal among LiSc estimators. This is made more precise in the following corollary, which also gives explicit asymptotic formulas for $r(\eta^2)$. The corollary follows immediately from Proposition 2 and Theorem 2.

Corollary 1. *For* $\rho \in [0, 1)$ *and* $\eta \geq 0$, *define the asymptotic risk functions*

$$R_0(\eta^2, \rho) = \frac{\eta^2 \rho}{\eta^2 + \rho} \text{ and } R_{>0}(\eta^2, \rho) = \frac{\eta^2 \rho}{\eta^2(1-\rho) + \rho}. \tag{10}$$

If $0 < \rho < 1$, *then*

$$\lim_{d/n \to \rho} \sup_{\eta \geq 0} \left| R_{>0}(\eta^2, d/n) - r(\eta^2) \right| = \lim_{d/n \to \rho} \sup_{\eta \geq 0} \sup_{V \in \Theta^d(\eta^2)} \left| R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} - r(\eta^2) \right| = 0.$$

Additionally,

$$\lim_{\substack{d/n \to 0 \\ d \to \infty}} \sup_{\eta \geq 0} \left| \frac{r(\eta^2)}{R_0(\eta^2, d/n)} - 1 \right| = \lim_{\substack{d/n \to 0 \\ d \to \infty}} \sup_{\eta \geq 0} \sup_{V \in \Theta^d(\eta^2)} \left| \frac{r(\eta^2)}{R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\}} - 1 \right| = 0.$$

Taken together, Theorem 1 and Corollary 1 provide exact formulas for the asymptotic behavior of $r(\eta^2)$ in any setting where $d \to \infty$. This is summarized in Table 1. We emphasize that if $\rho, \eta > 0$ are fixed, then the limiting LiSc risk $\lim_{d/n \to \rho} r(\eta^2) > 0$ is non-zero; on the other hand, $\lim_{d/n \to 0} \sup_{\eta \geq 0} r(\eta^2) = 0$.

TABLE 1
Asymptotics for the minimal LiSc risk

$d/n \to \rho$	Asymptotic approximation for $r(\eta^2)$
$\rho = 0$	$R_0(\eta^2, d/n)$
$\rho \in (0, 1)$	$R_{>0}(\eta^2, d/n)$
$\rho \geq 1$	η^2

The asymptotic risk formula $R_0(\eta^2, \rho) \sim r(\eta^2)$ in Corollary 1, which is valid when $d/n \to 0$, appears frequently in minimax analyses involving the Gaussian sequence model (Nussbaum, 1999; Pinsker, 1980). On the other hand, if $d/n \to \rho \in (0,1)$, then $r(\eta^2) \sim R_{>0}(\eta^2, d/n) > R_0(\eta^2, d/n)$. This reflects the increased difficulty in prediction problems where d/n is substantially larger than 0 and may be attributed to a degrees of freedom correction that accounts for the number of predictors in the linear model; in particular, the effect of this correction is non-vanishing when $d/n \to \rho \in (0,1)$.

4.2. Adaptive James-Stein estimators

The oracle James-Stein estimator $\hat{\boldsymbol{\beta}}_{js}(\eta^2)$ depends on the signal-to-noise ratio $\eta^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2$, which is typically unknown. If $d/n \to \rho < 1$, then the signal-to-noise ratio may be effectively estimated and it is reasonable to replace η^2 in $\hat{\boldsymbol{\beta}}_{js}(\eta^2)$ with an estimate. For $d < n$, define the estimator

$$\hat{\eta}^2_+ = \max\{\hat{\eta}^2, 0\} = \max\left\{\frac{||\mathbf{y}||^2}{n\hat{\sigma}^2} - 1, 0\right\}, \tag{11}$$

where $\hat{\eta}^2 = ||\mathbf{y}||^2/(n\hat{\sigma}^2) - 1$ was introduced in Theorem 1 (a). Note that if $V \in \Theta^d(\eta^2)$ and $n, n-d$ are large, then $n^{-1}||\mathbf{y}||^2 \approx \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} + \sigma^2$ and $\hat{\sigma}^2 \approx \sigma^2$, which suggests that $\hat{\eta}^2_+ \approx \eta^2$. Now define the adaptive James-Stein estimator

$$\check{\boldsymbol{\beta}}_{js} = \hat{\boldsymbol{\beta}}_{js}(\hat{\eta}^2_+).$$

Observe that $\check{\boldsymbol{\beta}}_{js}$ *adapts* to the unknown signal-to-noise ratio η^2. Furthermore, $\check{\boldsymbol{\beta}}_{js}$ is an LiSc estimator. The next result implies that if n is large and d/n is bounded below 1, then the predictive risk of the adaptive James-Stein estimator is almost as small as that of the oracle James-Stein estimator.

Theorem 3. *Suppose that* $0 < d/n < \rho_+ < 1$*, where* $\rho_+ \in \mathbb{R}$ *is a fixed constant. Then*

$$\sup_{V \in \Theta^d(\eta^2)} \left| R_V(\check{\boldsymbol{\beta}}_{js}) - R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} \right| = O\left\{ \left(\frac{d/n}{\eta^2 + d/n} \right) n^{-1/2} \right\}.$$

Theorem 3 is proved in Appendix A. It follows from Theorem 3 that if $d/n \to \rho \in [0,1)$, then the predictive risk of the adaptive James-Stein estimator converges uniformly to that of the oracle James-Stein estimator. Note that Theorem 3 is less informative when the signal-to-noise ratio is very small. Indeed, if $\eta^2 = O(n^{-1/2})$ and $d/n \to \rho \in [0,1)$, then $\sup_{V \in \Theta^d(\eta^2)} R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} = O(n^{-1/2})$ has the same magnitude as the upper bound in Theorem 3. A more refined analysis of the adaptive James-Stein estimator when the signal-to-noise ratio is small may be of interest, but is not pursued further here.

The following corollary is an immediate consequence of Theorems 2-3.

Corollary 2. *Suppose that* $\rho \in (0,1)$. *Then*

$$\lim_{d/n \to \rho} \sup_{\eta \geq 0} \sup_{V \in \Theta^d(\eta^2)} \left| R_V(\check{\boldsymbol{\beta}}_{js}) - r(\eta^2) \right| = 0.$$

Corollary 2 implies that if $d/n \to \rho \in (0,1)$, then the adaptive James-Stein estimator is asymptotically optimal for predictive risk among LiSc estimators.

5. Comparative analysis

5.1. OLS estimator

The predictive risk of the OLS estimator follows immediately from Proposition 2: if $d < n-1$, then

$$R_V(\hat{\boldsymbol{\beta}}_{ols}) = \frac{d}{n-d-1}. \tag{12}$$

Proposition 2 also implies that if $V \in \Theta^d(\eta^2)$ for some $\eta \geq 0$, then $R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} < R_V(\hat{\boldsymbol{\beta}}_{ols})$. On the other hand, as discussed in detail above, the oracle James-Stein estimator $\hat{\boldsymbol{\beta}}_{js}(\eta^2)$ is generally not implementable, because the signal-to-noise ratio $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2 = \sigma^2$ is typically unknown. The adaptive James-Stein estimator $\check{\boldsymbol{\beta}}$ does not depend on the signal-to-noise ratio and in Section 4.2 we argued that it is asymptotically equivalent to the oracle James-Stein estimator. The next result is valid in finite samples, for d, n sufficiently large, and relates the risk of the adaptive James-Stein estimator to that of the OLS estimator.

Theorem 4. *Suppose that* $0 < \rho_- \leq d/n \leq \rho_+ < 1$ *for some fixed constants* $\rho_-, \rho_+ \in \mathbb{R}$. *If* d, n *are sufficiently large, then*

$$R_V(\check{\boldsymbol{\beta}}_{js}) < R_V(\hat{\boldsymbol{\beta}}_{ols}) \tag{13}$$

for all $V \in PD(d+1)$.

Theorem 4 follows directly from Theorem 2, Theorem 3, and (12). Theorem 4 implies that $\check{\boldsymbol{\beta}}_{js}$ is minimax over the entire parameter space $PD(d+1)$, when d, n are sufficiently large. We refer to Theorem 4 as a "semi-finite sample" result, because it addresses finite sample properties of $\check{\boldsymbol{\beta}}_{js}$, but we are unable to specify precisely how large d, n must be in order for (13) to hold. Theorem 4 may be contrasted with more classical finite sample results on James-Stein estimators for the normal means problem (James and Stein, 1961) and linear models (Baranchik, 1973), which imply that certain James-Stein estimators are minimax under explicit conditions on the dimension; for instance, Baranchik (1973) shows that a different James-Stein estimator for $\boldsymbol{\beta}$ (which is discussed in more detail in Section 5.3) is minimax whenever $d > 2$ and $n - d > 1$. Similar results may be available for the adaptive James-Stein estimator $\check{\boldsymbol{\beta}}_{js}$; however, this paper is more focused on high-dimensional optimality properties of $\check{\boldsymbol{\beta}}_{js}$, like those discussed in Section 4, and alternative techniques are likely required to obtain more detailed finite sample results.

5.2. Ridge regression

Ridge regression (Hoerl and Kennard, 1970; Tikhonov, 1943) is another widely studied non-sparse estimator. For a positive definite matrix $\Lambda \in PD(d)$, define the generalized ridge estimator

$$\hat{\boldsymbol{\beta}}_r(\Lambda) = (X^T X + n\Lambda)^{-1} X^T \mathbf{y}.$$

Note that $\hat{\boldsymbol{\beta}}_r(\Lambda)$ is defined for all d, n. While generalized ridge estimators have been studied for many classes of Λ, by far the most common is $\Lambda = \lambda I$, where $\lambda > 0$ is a scalar shrinkage factor subject to further specification. Note, however, that for fixed $\lambda > 0$, $\hat{\boldsymbol{\beta}}_r(\lambda I)$ is *not* LiSc. Furthermore, the following result suggests that $\hat{\boldsymbol{\beta}}_r(\lambda I)$ has significant drawbacks (in a minimax sense) when its performance is evaluated over linear models with fixed signal-to-noise ratio.

Proposition 3. *Suppose that $\eta \geq 0$. Then*

$$\inf_{\lambda > 0} \sup_{V \in \Theta^d(\eta^2)} R_V\{\hat{\boldsymbol{\beta}}_r(\lambda I)\} \geq \sup_{V \in \Theta^d(\eta^2)} R_V(\hat{\boldsymbol{\beta}}_{null}) = \eta^2.$$

Proposition 3 is proved in Appendix A. It implies that the ridge estimator $\hat{\boldsymbol{\beta}}_r(\lambda I)$'s worst-case out-of-sample prediction error is at least as bad as that of the trivial estimator $\hat{\boldsymbol{\beta}}_{null} = 0$, over linear models with fixed signal-to-noise ratio.

As an alternative to $\hat{\boldsymbol{\beta}}_r(\lambda I)$, we consider a generalized ridge estimator that depends on the predictor covariance matrix $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$. For $V \in \Theta^d(\eta^2)$ given by (4), define the oracle ridge estimator

$$\hat{\boldsymbol{\beta}}_r\{d/(n\eta^2)\Sigma\} = (X^T X + d/\eta^2 \Sigma)^{-1} X^T \mathbf{y}. \tag{14}$$

To motivate this estimator, we note that in (Dicker, 2013), the author considered ridge regression in high dimensional linear models where $\mathrm{Cov}(\mathbf{x}_i) = I$ and $\sigma^2 = 1$; the oracle ridge estimator (14) corresponds to oracle estimators derived in (Dicker, 2013), after transforming the data via $(\mathbf{y}, X) \mapsto (\mathbf{y}, X\Sigma^{-1/2})$. It follows that $\hat{\boldsymbol{\beta}}_r\{d/(n\eta^2)\Sigma\}$ shares many optimality properties with the ridge estimators studied in (Dicker, 2013). Adaptive ridge estimators may be derived by replacing η^2 and Σ in (14) with estimates, $\hat{\eta}^2$ and $\hat{\Sigma}$; if $\hat{\eta}^2$ is consistent for η^2 and $\hat{\Sigma}$ is operator norm-consistent for Σ, then the associated adaptive ridge estimator is typically asymptotically equivalent to the oracle ridge estimator.

The oracle ridge estimator (14) satisfies an equivariance property for estimators depending on $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ that extends the LiSc property given in Definition 1. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X, \Sigma)$ be an estimator for $\boldsymbol{\beta}$ that may depend on $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$. Then $\hat{\boldsymbol{\beta}}$ is LiSc if

$$A^{-1}\hat{\boldsymbol{\beta}}(\mathbf{y}, X, \Sigma) = \hat{\boldsymbol{\beta}}(t\mathbf{y}, tXA, t^2 A^T \Sigma A) \tag{15}$$

for all $d \times d$ invertible matrices $A \in GL(d)$ and all $t > 0$. Thus, an LiSc estimator's dependence on Σ must respect linear transformations of the predictors

$\mathbf{x}_i \sim N(0, \mathbf{x}_i)$. Clearly, if $\hat{\boldsymbol{\beta}}$ does not depend on $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$, then (15) reduces to the LiSc property given in Definition 1. Furthermore, the oracle ridge estimator (14) is LiSc. We emphasize that the oracle ridge estimator is LiSc, even for $d > n$; by contrast, Theorem 1 (c) implies that if $d > n$, then the null estimator is the *only* LiSc estimator that does not depend on Σ.

Some of the basic risk properties of $\hat{\boldsymbol{\beta}}_r\{d/(n\eta^2)\Sigma\}$ are given in the following proposition.

Proposition 4. *Suppose that $\eta \geq 0$ is fixed.*

(a) [Finite sample predictive risk] *If $V \in \Theta^d(\eta^2)$, then*

$$R_V\left[\hat{\boldsymbol{\beta}}_r\left\{d/(n\eta^2)\Sigma\right\}\right] = E_I\left[\mathrm{tr}\left\{(X^TX + d/\eta^2 I)^{-1}\right\}\right].$$

(b) [Asymptotic predictive risk] *Suppose that $\rho \in (0, \infty)$ and define*

$$R_r(\eta^2, \rho) = \frac{1}{2\rho}\left[\eta^2(\rho - 1) - \rho + \sqrt{\{\eta^2(\rho - 1) - \rho\}^2 + 4\eta^2\rho^2}\right]. \tag{16}$$

Then

$$\lim_{d/n \to \rho} \sup_{V \in \Theta^d(\eta^2)} R_V\left[\hat{\boldsymbol{\beta}}_r\left\{d/(n\eta^2)\Sigma\right\}\right] = R_r(\eta^2, \rho).$$

Proposition 4 follows immediately from Proposition 1 and Corollary 1 in (Dicker, 2013). Proposition 4 (a) gives a simplified expression for the predictive risk of the oracle ridge estimator; in particular, it implies that $R_V[\hat{\boldsymbol{\beta}}_r\{d/(n\eta^2)\Sigma\}]$ is completely determined by the signal-to-noise ratio $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}/\sigma^2 = \eta^2$. Proposition 4 (b) gives a closed-form expression for the asymptotic predictive risk of the oracle ridge estimator that is valid when $d/n \to \rho \in (0, \infty)$.

It is evident from Proposition 4 that the oracle ridge estimator has smaller risk than $\hat{\boldsymbol{\beta}}_{null} = 0$, even when $d > n$; that is, if $V \in \Theta^d(\eta^2)$, then

$$R_V\left[\hat{\boldsymbol{\beta}}_r\left\{d/(n\eta^2)\Sigma\right\}\right] \leq R_V(\hat{\boldsymbol{\beta}}_{null})$$

with equality if and only if $\eta^2 = 0$. Since the oracle ridge estimator is LiSc, it follows that if $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is known, then non-trivial equivariant out-of-sample prediction may be possible when $d > n$; on the other hand, Theorem 1 (c) implies that this is impossible if $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is unknown.

Proposition 4 also yields detailed information about the asymptotic predictive risk of the oracle ridge estimator (14), which is useful for comparing its performance to that of the oracle James-Stein estimator. Corollary 1 and Proposition 4 (a) imply that if $d/n \to 0$, then

$$\sup_{V \in \Theta^d(\eta^2)} R_V\left[\hat{\boldsymbol{\beta}}_r\left\{d/(n\eta^2)\Sigma\right\}\right] \sim \sup_{V \in \Theta^d(\eta^2)} R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} \sim R_0(\eta^2, d/n),$$

where $R_0(\eta^2, d/n)$ is given in Corollary 1. Thus, if $d/n \to 0$, then the oracle ridge and James-Stein estimators are asymptotically equivalent. On the other

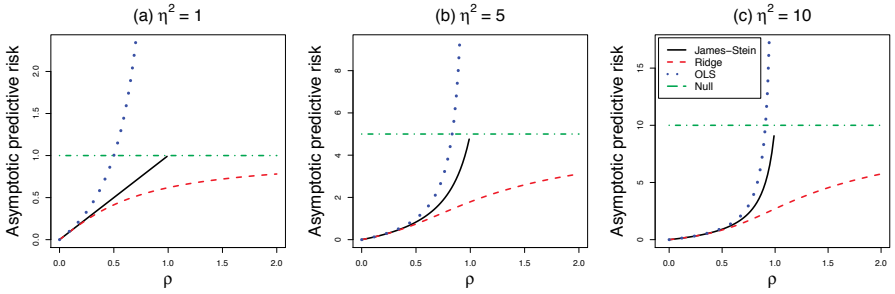

FIG 1. *Asymptotic predictive risk versus ρ for the oracle James-Stein estimator ($R_{>0}(\eta^2, \rho)$, defined in Corollary 2), the oracle ridge estimator ($R_r(\eta^2, \rho)$, defined in Proposition 6), the OLS estimator ($R_{ols}(\eta^2, \rho)$, defined in (17)), and the null estimator ($R_{null}(\eta^2, \rho)$, defined in (17)) for various values of the signal-to-noise ratio $\eta^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}/\sigma^2$: (a) $\eta^2 = 1$, (b) $\eta^2 = 5$, and (c) $\eta^2 = 10$. The James-Stein and OLS estimators are undefined when $d > n$; thus, their predictive risk plots stop at $\rho = 1$.*

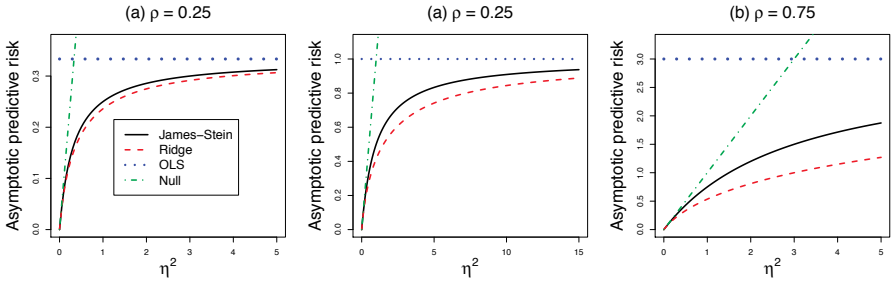

FIG 2. *Asymptotic predictive risk versus ρ for the oracle James-Stein estimator ($R_{>0}(\eta^2, \rho)$, defined in Corollary 2), the oracle ridge estimator ($R_r(\eta^2, \rho)$, defined in Proposition 6), the OLS estimator ($R_{ols}(\eta^2, \rho)$, defined in (17)), and the null estimator ($R_{null}(\eta^2, \rho)$, defined in (17)) for various values of ρ: (a) $\rho = 0.25$, (b) $\rho = 0.50$, and (c) $\rho = 0.75$.*

hand, one easily checks that if $\rho \in (0, \infty)$, then $R_r(\eta^2, \rho) \leq R_{>0}(\eta^2, \rho)$ with equality if and only if $\eta^2 = 0$. Thus, if $d/n \to \rho \in (0, \infty)$, then the ridge estimator outperforms the James-Stein estimator in terms of predictive risk. This should not be too surprising, because the ridge estimator utilizes knowledge of $\text{Cov}(\mathbf{x}_i) = \Sigma$, while the James-Stein estimator does not. Figures 1-2 contain plots of the asymptotic predictive risk of the oracle ridge and James-Stein estimator, for $d/n \to \rho \in (0, \infty)$. The asymptotic predictive risk of the OLS estimator and the null estimator are also plotted in Figures 1-2; it is easily seen that if $d/n \to \rho \in (0, \infty)$, then the asymptotic predictive risk of $\hat{\boldsymbol{\beta}}_{ols}$ and $\hat{\boldsymbol{\beta}}_{null}$ is given by

$$R_{ols}(\eta^2, \rho) = \frac{\rho}{1 - \rho} \quad (0 \leq \rho < 1) \quad \text{and} \quad R_{null}(\eta^2, \rho) = \eta^2, \tag{17}$$

respectively (the asymptotic risk for the OLS estimator $R_{ols}(\eta^2, \rho)$ follows directly from (12)).

To conclude this subsection, we give a simple result which implies that the oracle ridge estimator dominates the oracle James-Stein estimator in finite samples. Again, this is not surprising because the ridge estimator leverages knowledge of $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$, while the James-Stein estimator does not.

Proposition 5. *Suppose that $d < n-1$, $\eta \geq 0$, and $V \in \Theta^d(\eta)$. Then*

$$R_V\left[\hat{\boldsymbol{\beta}}_r\left\{d/(n\eta^2)\Sigma\right\}\right] \leq R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\}$$

with equality if and only if $\eta = 0$.

Proposition 5 follows from Jensen's inequality and is proved in Appendix A.

5.3. Other James-Stein estimators

Other James-Stein type estimators for $\boldsymbol{\beta}$ have been previously studied in the literature (Baranchik, 1973; Brown, 1990; Copas, 1983; Oman, 1984; Stein, 1960; Takada, 1979). Much of the previous work on James-Stein estimators for $\boldsymbol{\beta}$ focuses on identifying situations where the specified estimators have uniformly smaller predictive risk than the OLS estimator in finite samples. To our knowledge, the asymptotic risk of other James-Stein estimators for $\boldsymbol{\beta}$ in high-dimensional linear models (with $d/n \to \rho \in (0,1)$) has received relatively little attention. In this section, we derive the asymptotic predictive risk of a James-Stein estimator for $\boldsymbol{\beta}$ studied by Baranchik (1973). For $d < n$ and constant $c > 0$, this estimator is defined by

$$\hat{\boldsymbol{\beta}}_{bar}(c) = \left\{1 - c\frac{||\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ols}||^2}{||X\hat{\boldsymbol{\beta}}_{ols}||^2}\right\}\hat{\boldsymbol{\beta}}_{ols} = \hat{\boldsymbol{\beta}}_{js}\{\hat{t}_{bar}(c)\},$$

where

$$\hat{t}_{bar}(c) = \frac{d}{n-d-1}\left(\frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{c||\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ols}||^2} - 1\right).$$

Baranchik (1973) proved that if $0 < c < 2(d-2)/(n-d+2)$, $d \geq 3$, and $n \geq d+2$, then $R_V\{\hat{\boldsymbol{\beta}}_{bar}(c)\} < R_V(\hat{\boldsymbol{\beta}}_{ols})$. Other previously studied James-Stein estimators share strong similarities with $\hat{\boldsymbol{\beta}}_{bar}(c)$. For instance, Copas (1983) considers precisely $\hat{\boldsymbol{\beta}}_{bar}(c)$ and provides arguments for using various specific values of c, while $\hat{\boldsymbol{\beta}}_{bar}(c)$ serves as a motivating example for a more general class of James-Stein estimators proposed by Takada (1979). Many of these estimators can be analyzed using techniques similar to those found in this subsection, and throughout the paper.

Below, we show that $\hat{\boldsymbol{\beta}}_{bar}(c)$ is generally sub-optimal, in terms of predictive risk, and that it is out-performed (asymptotically) by the adaptive James-Stein estimator $\check{\boldsymbol{\beta}}_{js}$ defined in Section 4.2. On the other hand, we also show that $\hat{\boldsymbol{\beta}}_{bar}(c)$ is asymptotically optimal for another, closely related loss function. The main idea behind our asymptotic analysis is that if $V \in \Theta(\eta^2)$, $d/n \approx \rho \in (0,1)$,

and n is large, then $||\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ols}||^2 \approx n(1-\rho)\sigma^2$ and $||\mathbf{y}||^2 \approx n(\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \sigma^2)$. Thus,

$$\hat{t}_{bar}(c) \approx t_{bar}(c) \triangleq \frac{\rho}{1-\rho}\left\{\frac{\eta^2 + \rho - c(1-\rho)}{c(1-\rho)}\right\}$$

and

$$\hat{\boldsymbol{\beta}}_{bar}(c) \approx \hat{\boldsymbol{\beta}}_{js}\{t_{bar}(c)\}. \tag{18}$$

By Proposition 2, the risk of the James-Stein estimator $\hat{\boldsymbol{\beta}}_{js}(t)$ is minimized when $t = \eta^2$; moreover, if $t \neq \eta^2$, then $R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta^2)\} < R_V\{\hat{\boldsymbol{\beta}}_{js}(t)\}$ and $\hat{\boldsymbol{\beta}}_{js}(t)$ is suboptimal. Since, in general, $t_{bar}(c) \neq \eta^2$, (18) suggests that $\hat{\boldsymbol{\beta}}_{bar}(c)$ is suboptimal among James-Stein estimators. Observe that while the equality $t_{bar}(c) = \eta^2$ may hold for some specific values of c, ρ, and η^2, in order for it to hold in general, the constant c from $\hat{\boldsymbol{\beta}}_{bar}(c)$ must vary with ρ and η^2.

Some of the ideas from the previous discussion are made more rigorous in the next proposition. A detailed proof is omitted; however, part (a) is a straightforward calculation, part (b) may be proved similar to Theorem 3, and part (c) follows directly from part (b), Corollary 1, and Theorem 3.

Proposition 6. *(a) Suppose that $\rho \in (0,1)$ and let*

$$R(\eta^2, \rho; t) = \left(\frac{t}{t + \frac{\rho}{1-\rho}}\right)^2 \frac{\rho}{1-\rho} + \left(\frac{\frac{\rho}{1-\rho}}{t + \frac{\rho}{1-\rho}}\right)^2 \eta^2$$

denote the asymptotic predictive risk of the James-Stein estimator $\hat{\boldsymbol{\beta}}_{js}(t)$ as $d/n \to \rho$. Then

$$R_{>0}(\eta^2, \rho) = R(\eta^2, \rho; \eta^2) \leq R\{\eta^2, \rho; t_{bar}(c)\}, \tag{19}$$

where $R_{>0}(\eta^2, \rho)$ is the asymptotic risk of the oracle James-Stein estimator defined in Corollary 2. Furthermore, equality holds in (19) if and only if

$$c = \frac{\eta^2\rho + \rho^2}{\eta^2(1-\rho)^2 + \rho(1-\rho)}.$$

(b) Suppose that $0 < \rho^- \leq d/n \leq \rho^+ < 1$ for some fixed constants $\rho^-, \rho^+ \in \mathbb{R}$ and that c is a positive constant satisfying $0 < c < 2\rho_-/(1-\rho_-)$ for all n and d. Let $\rho = d/n$. Then

$$R_V\{\hat{\boldsymbol{\beta}}_{bar}(c)\} = R_V\left[\hat{\boldsymbol{\beta}}_{js}\{t_{bar}(c)\}\right] + O\left\{\frac{1}{(\eta^2+1)n^{1/2}}\right\}.$$

(c) Under the assumptions of part (b),

$$\begin{aligned} R_V\{\hat{\boldsymbol{\beta}}_{bar}(c)\} - R_V(\check{\boldsymbol{\beta}}_{js}) &= R\{\rho, \eta^2; t_{bar}(c)\} - R_{>0}(\eta^2, \rho) \\ &\quad + O\left\{\frac{1}{(\eta^2+1)n^{1/2}}\right\}, \end{aligned}$$

where $\check{\boldsymbol{\beta}}_{js}$ is the adaptive James-Stein estimator from Section 4.2.

Part (a) of Proposition 6 addresses suboptimality of $\hat{\boldsymbol{\beta}}_{js}\{t_{bar}(c)\}$ and part (b) provides justification for (18). Proposition 6 (c) implies that the predictive risk of the adaptive James-Stein estimator $\check{\boldsymbol{\beta}}_{js}$ is asymptotically smaller than that of $\hat{\boldsymbol{\beta}}_{bar}(c)$.

It follows from Proposition 6 that $\hat{\boldsymbol{\beta}}_{bar}(c)$ is suboptimal in terms of predictive risk, even among the class of James-Stein estimators, $\hat{\boldsymbol{\beta}}_{js}(t)$. This naturally leads to the question: are there other circumstances under which Baranchik's estimator $\hat{\boldsymbol{\beta}}_{bar}(c)$ is asymptotically optimal among James-Stein estimators? The answer is affirmative. Consider a prediction problem where the predictors $\mathbf{x}_{new}$ associated with future outcomes y_{new} are required to be drawn from $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. If we assume that $P\{\mathbf{x}_{new} = \mathbf{x}_i | X\} = n^{-1}$, $i = 1, \ldots, n$, then a reasonable measure of predictive risk is

$$\tilde{R}_V(\hat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2 n} E_V ||X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2.$$

Now let $t^*_{bar} = n\eta^2/(n - d - 1)$. It is straightforward to check that

$$\tilde{R}_V\{\hat{\boldsymbol{\beta}}_{js}(t^*_{bar})\} = \inf_{t \in [0,\infty]} \tilde{R}_V\{\hat{\boldsymbol{\beta}}_{js}(t)\}$$

and, if n is large, then $\hat{\boldsymbol{\beta}}_{bar}\{d/(n-d)\} \approx \hat{\boldsymbol{\beta}}_{js}(t^*_{bar})$. Ultimately, one can show that if $d/n \to \rho \in (0, 1)$, then $\hat{\boldsymbol{\beta}}_{bar}\{d/(n-d)\}$ is asymptotically optimal among James-Stein estimators, with respect to the risk function $\tilde{R}_V(\cdot)$. In fact, Copas (1983) reaches a similar conclusion – that one should take $c \approx d/(n-d)$ in $\hat{\boldsymbol{\beta}}_{bar}(c)$ – by essentially studying the risk function $\tilde{R}_V(\cdot)$. However, Copas (1983) does not take an asymptotic approach, nor does Copas meaningfully distinguish between the risk functions $R_V(\cdot)$ and $\tilde{R}_V(\cdot)$. Indeed, Copas asserts that differences between the two risk functions are "unimportant if n is large" (p. 314 of (Copas, 1983)). This is true if n is large and d is small; however, the results in this section imply that these differences are significant when both n and d are large.

6. Discussion

6.1. Distributional assumptions

The normality condition (2) is restrictive. The extent to which this condition is necessary for the results in this paper is somewhat unclear; working to relax (2) may be an interesting area for future research. In this section, we discuss some of the issues that may arise in pursuing such work.

If the data are non-Gaussian, then the exact risk formula for James-Stein estimators given in Proposition 2 will not hold in general (among other things, Proposition 2 relies on the fact that $E_I\{\text{tr}(X^T X)^{-1}\} = d/(n - d - 1)$). Furthermore, in the absence of normality, it may be more challenging to obtain exact decision theoretic results for LiSc estimators, such as Theorem 1 (a)-(b), which rely on orthogonal invariance of the multivariate normal distribution

(note, however, that Theorem 1 (c) continues to hold regardless of distributional assumptions: $\hat{\boldsymbol{\beta}}_{null} = 0$ is the only LiSc estimator when $d > n$).

The challenges discussed in the previous paragraph may be complemented by more encouraging observations. For instance, Stein-type estimators are known to have desirable finite sample properties in related problems with non-Gaussian data (see, for example, the review article (Brandwein and Strawderman, 1990) on estimating a location parameter in the presence of orthogonally invariant noise); these results may be relevant for generalizing the results in this paper to settings where the data are non-normal. Additionally, it seems reasonable to expect that many of the asymptotic results in this paper (or close variants) will continue to hold under weaker distributional assumptions – even in settings where the underlying distributions are not orthogonally invariant. Basic numerical experiments conducted by the author seem to support this hypothesis when the entries of X are binary random variables (detailed results not reported here). Existing theoretical work on high-dimensional data analysis with non-Gaussian design matrices may also be useful for establishing extensions in this direction, e.g. (Bunea et al., 2007a).

6.2. Practical implications

We have argued that LiSc estimators are a reasonable class of estimators for settings where little is known about how the given predictor basis relates to the outcome of interest. In these settings, if $d < n$ and no reliable estimate of $\mathrm{Cov}(\mathbf{x}_i)$ is available, then the results in this paper suggest that James-Stein estimators may be an effective option for out-of-sample prediction; if $\mathrm{Cov}(\mathbf{x}_i)$ is known (or if a norm-consistent estimator is available), then results in Section 5.2 and (Dicker, 2013) imply that ridge regression may be more appropriate. On the other hand, if $\boldsymbol{\beta}$ is known to be sparse (i.e. if the outcome has a sparse representation in the given predictor basis) or some other significant prior information about $\boldsymbol{\beta}$ is available, then sparse methods, such as lasso, or Bayesian methods may be indicated (it is worth pointing out that $\hat{\boldsymbol{\beta}}_{js}(t)$ is a Bayes estimator under the prior distribution $\boldsymbol{\beta}|X \sim N\{0, \nu^2(X^TX)^{-1}\}$, where $\nu^2 = t(n-d-1)\sigma^2/d$).

If $d \geq n$, then Theorem 1 (b)-(c) imply that $\hat{\boldsymbol{\beta}}_{null} = 0$ is the optimal LiSc estimator. Thus, one can argue that if $d \geq n$, then requiring an estimator to be LiSc is "asking too much." On the other hand, given the high-level discussion of LiSc estimators in the previous paragraph and in Section 3, an alternative interpretation of Theorem 1 (b)-(c) is as follows: if $d \geq n$, $\mathrm{Cov}(\mathbf{x}_i)$ is unknown, and little is known about how the predictor basis relates to the outcome, then non-trivial prediction may be impossible and, consequently, more information is needed for effective out-of-sample prediction. This interpretation may guide one's view towards identifying and understanding additional information about the model and the data that may help to improve performance in settings where $d \geq n$.

Various types of information about the model (1) may potentially be leveraged to develop better prediction methods. The discussion of ridge regression in

Section 5.2 implies that if $\mathrm{Cov}(\mathbf{x}_i)$ is known, then an equivariant version of ridge regression (14) may perform well (significantly better than $\hat{\boldsymbol{\beta}}_{null}$) when $d \geq n$. However, to obtain more substantial improvements in out-of-sample prediction error, it appears that additional information about $\boldsymbol{\beta}$ (such as sparsity) must be utilized. Indeed, the predictive risk of ridge regression is roughly of order d/n; if $\boldsymbol{\beta}$ is sparse, then the risk of lasso may be of order $\log(d)/n$ (Bunea et al., 2007b). Slightly recasting these observations, we conclude that while ridge regression outperforms $\hat{\boldsymbol{\beta}}_{null} = 0$ when $d/n \to \rho \in (0, \infty)$, additional information about $\boldsymbol{\beta}$ must be utilized in order to obtain vanishing risk in these asymptotic settings and, a fortiori, in settings where $d/n \to \infty$.

Acknowledgements

The author thanks Patrick O. Perry for valuable comments on an earlier version of this paper. The author is grateful to the Editor, Associate Editor, and Referees for many comments that helped to improve the paper.

Appendices

Appendix A

Proof of Proposition 1. Suppose that $\hat{\boldsymbol{\beta}}$ is linearly equivariant. To prove part (a), observe that

$$\begin{aligned}
\sigma^2 R_V(\hat{\boldsymbol{\beta}}) &= E_V\left[\{\hat{\boldsymbol{\beta}}(y, X, \Sigma) - \boldsymbol{\beta}\}^T \Sigma \{\hat{\boldsymbol{\beta}}(y, X, \Sigma) - \boldsymbol{\beta}\}\right] \\
&= E_V\left\{||\hat{\boldsymbol{\beta}}(y, X\Sigma^{-1/2}, I) - \Sigma^{1/2}\boldsymbol{\beta}||^2\right\} \\
&= E_{V_0}\left\{||\hat{\boldsymbol{\beta}}(y, X, I) - \boldsymbol{\beta}(V_0)||^2\right\} \\
&= \sigma^2 R_{V_0}(\hat{\boldsymbol{\beta}}),
\end{aligned}$$

where $\boldsymbol{\beta}(V_0) = \Sigma^{1/2}\boldsymbol{\beta}$, and we have used linear equivariance of $\hat{\boldsymbol{\beta}}$, along with the fact that if $\mathbf{x}_i \sim N(0, \Sigma)$, then $\Sigma^{-1/2}\mathbf{x}_i \sim N(0, I)$.

Now suppose that $\hat{\boldsymbol{\beta}}$ is LiSc, let $\mathbf{u} \in \mathbb{R}^d$ be a unit vector, and let U be a $d \times d$ orthogonal matrix such that $U\Sigma^{1/2}\boldsymbol{\beta}/\sigma = \eta\mathbf{u} \triangleq \boldsymbol{\beta}(V_{\mathbf{u}})$. Then

$$\begin{aligned}
R_V(\hat{\boldsymbol{\beta}}) &= \sigma^{-2} E_{V_0}\left(||\hat{\boldsymbol{\beta}} - \Sigma^{1/2}\boldsymbol{\beta}||^2\right) \\
&= E_{V_0}\left\{||\sigma^{-1}U\hat{\boldsymbol{\beta}}(y, X, I) - \eta\mathbf{u}||^2\right\} \\
&= E_{V_0}\left\{||\hat{\boldsymbol{\beta}}(XU^T\eta\mathbf{u} + \boldsymbol{\epsilon}/\sigma, XU^T, I) - \eta\mathbf{u}||^2\right\} \\
&= E_{V_{\mathbf{u}}}\left\{||\hat{\boldsymbol{\beta}}(y, X, I) - \boldsymbol{\beta}(V_{\mathbf{u}})||^2\right\} \\
&= R_{V_{\mathbf{u}}}(\hat{\boldsymbol{\beta}}).
\end{aligned}$$

□

Proof of Theorem 1. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X)$ be an LiSc estimator. Let $O(n)$ denote the group of $n \times n$ orthogonal matrices. Since $\hat{\boldsymbol{\beta}}(\mathbf{y}, X)$ and $\hat{\boldsymbol{\beta}}(U\mathbf{y}, UX)$ have the same distribution for all $U \in O(n)$, for parts (a)-(b) of the theorem we may assume without loss of generality that

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}(U\mathbf{y}, UX) \text{ for all } U \in O(n). \tag{20}$$

To prove part (a), suppose that $d < n$. Then

$$\hat{\boldsymbol{\beta}} = h(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols}$$

for some measurable function $h : \mathbb{R} \to \mathbb{R}$. Now suppose that $V \in \Theta^d(\eta^2)$. Then,

$$\begin{aligned} R_V(\hat{\boldsymbol{\beta}}) &= \sigma^{-2}E_V\left[\{h(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\}^T\Sigma\{h(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\}\right] \\ &= E\left\{h^2(\hat{\eta}^2)\sigma^{-2}E_V(\hat{\boldsymbol{\beta}}_{ols}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}|\hat{\eta}^2)\right\} \\ &\quad -2\sigma^{-2}E\left\{h(\hat{\eta}^2)E_V(\boldsymbol{\beta}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}|\hat{\eta}^2)\right\} + \eta^2. \end{aligned} \tag{21}$$

Taking $h(\hat{\eta}^2) = h_{opt}(\hat{\eta}^2)$, where $h_{opt}(\hat{\eta}^2)$ is given in (7), minimizes the integrand in (21). Since $\hat{\eta}^2(\mathbf{y}, X) = \hat{\eta}^2(t\mathbf{y}, tXA)$ for all $t > 0$ and invertible $d \times d$ matrices A, it follows as in the proof of Proposition 1 that $\sigma^{-2}E_V(\hat{\boldsymbol{\beta}}_{ols}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}|\hat{\eta}^2)$ and $\sigma^{-2}h(\hat{\eta}^2)E_V(\boldsymbol{\beta}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}|\hat{\eta}^2)$ are constant over $V \in \Theta^d(\eta^2)$. Thus, $h_{opt}(\hat{\eta}^2)$ depends only on η^2, d, and n. Part (a) of Theorem 1 follows.

To prove part (b) of the theorem, suppose that $d = n$. Then $\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = X^{-1}\hat{\boldsymbol{\beta}}(\mathbf{y}, I)$. Furthermore, if $U \in O(d)$, then

$$\hat{\boldsymbol{\beta}}(U\mathbf{y}, I) = \hat{\boldsymbol{\beta}}(\mathbf{y}, U^T) = U\hat{\boldsymbol{\beta}}(\mathbf{y}, I).$$

It follows that $\hat{\boldsymbol{\beta}}(\mathbf{y}, I) = h(||\mathbf{y}||^2)\mathbf{y}$ for some function $h : \mathbb{R} \to \mathbb{R}$ and

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = h(||\mathbf{y}||^2)X^{-1}\mathbf{y}.$$

Since $\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}(t\mathbf{y}, tX)$ for all $t > 0$, it follows that $h(||\mathbf{y}||^2) = h(t||\mathbf{y}||^2)$. Thus, $h(||\mathbf{y}||^2) = c$ is constant. Part (b) of the theorem follows because $E_I[\text{tr}\{(X^TX)^{-1}\}] = \infty$.

Finally, to prove part (c), we no longer require (20), but we assume that $d > n$. Let $X^T = Q_1S^T$ be the QR decomposition of X^T, where Q_1 is a $d \times n$ matrix with orthonormal columns and S is an $n \times n$ lower triangular matrix. Let

$$A = (Q_1 \;\; Q_2)\begin{pmatrix} S^{-1} & 0 \\ C & D \end{pmatrix} = Q_1(S^{-1} \;\; 0) + Q_2(C \;\; D),$$

where $(Q_1 \;\; Q_2) \in O(d)$, C is a $(d-n) \times n$ matrix, and $D \in GL(d-n)$. Then

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = A\hat{\boldsymbol{\beta}}(\mathbf{y}, XA) = Q_1(S^{-1} \;\; 0)\hat{\boldsymbol{\beta}}(\mathbf{y}, (I \;\; 0)) + Q_2(C \;\; D)\hat{\boldsymbol{\beta}}(\mathbf{y}, (I \;\; 0)).$$

Since the above equality must hold for any $(d-n) \times n$ matrix C and any $D \in GL(d-n)$, it follows that $\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = 0$. Part (c) follows immediately. □

Proof of Proposition 2. Note that $(X^TX)^{-1}$ follows an inverse Wishart distribution and $E_I(X^TX)^{-1} = (n-d-1)^{-1}I$ (see Chapter 3 of (Muirhead, 1982), for instance). It follows that if $V \in \Theta^d(\eta^2)$, then

$$\begin{aligned} R_V\{\hat{\boldsymbol{\beta}}_{js}(t)\} &= \left\{\frac{t}{t+d/(n-d-1)}\right\}^2 E_I \mathrm{tr}\{(X^TX)^{-1}\} \\ &\quad + \left\{\frac{d/(n-d-1)}{t+d/(n-d-1)}\right\}^2 \eta^2 \\ &= \left\{\frac{t}{t+d/(n-d-1)}\right\}^2 \frac{d}{n-d-1} + \left\{\frac{d/(n-d-1)}{t+d/(n-d-1)}\right\}^2 \eta^2. \end{aligned}$$

Thus, (8). The rest of the proposition follows by basic calculus. □

Proof of Theorem 2. Suppose that $V \in \Theta^d(\eta^2)$ and let $\hat{\boldsymbol{\beta}}_{opt} = h_{opt}(\hat{\eta}^2)\hat{\boldsymbol{\beta}}_{ols}$ be the optimal LiSc estimator derived in Theorem 1 (a). Then $r(\eta^2) = R_V(\hat{\boldsymbol{\beta}}_{opt})$. To prove the theorem, we bound $|R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta)\} - R_V\{\hat{\boldsymbol{\beta}}_{opt}(\eta)\}|$. Since $R_V\{\hat{\boldsymbol{\beta}}_{opt}(\eta)\} \le R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta)\}$, we have the following decomposition:

$$0 \le R_V\{\hat{\boldsymbol{\beta}}_{js}(\eta)\} - R_V\{\hat{\boldsymbol{\beta}}_{mm}(\eta)\} = E_V(L_1) + E_V(L_2) + 2E_V(L_3), \tag{22}$$

where

$$\begin{aligned} L_1 &= \frac{1}{\sigma^2}\left[h_{opt}^2(\hat{\eta}^2) - \left\{\frac{\eta^2}{\eta^2+d/(n-d-1)}\right\}^2\right] ||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol{\epsilon}||^2 \\ L_2 &= \eta^2\left[\{1-h_{opt}(\hat{\eta}^2)\}^2 - \left\{\frac{d/(n-d-1)}{\eta^2+d/(n-d-1)}\right\}^2\right] \\ L_3 &= \frac{1}{\sigma^2}h_{opt}(\hat{\eta}^2)\{1-h_{opt}(\hat{\eta}^2)\}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}. \end{aligned}$$

We bound $E_V(L_1)$, $E_V(L_2)$, and $E_V(L_3)$ separately.

To bound $E_V(L_1)$ and $E_V(L_2)$, we have

$$\begin{aligned} E_V|L_1| &\le \left\{E_V\left|h_{opt}(\hat{\eta}^2) - \frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^3\right\}^{1/3} \\ &\quad \cdot \left\{E_V\left|h_{opt}(\hat{\eta}^2) + \frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^3\right\}^{1/3} \\ &\quad \cdot \left\{\frac{1}{\sigma^6}E_V||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol{\epsilon}||^6\right\}^{1/3}, \\ E_V|L_2| &\le \left\{E_V\left|h_{opt}(\hat{\eta}^2) - \frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^2\right\}^{1/2} \\ &\quad \cdot \eta^2\left\{E_V\left|h_{opt}(\hat{\eta}^2) - 1 - \frac{d/(n-d-1)}{\eta^2+d/(n-d-1)}\right|^2\right\}^{1/2}. \end{aligned}$$

Thus, Lemma B5 implies that

$$\sup_{V\in\Theta^d(\eta^2)} E_V|L_1|,\ \sup_{V\in\Theta^d(\eta^2)} E_V|L_2| = O\left(\frac{\eta^2 d/n}{\eta^2+d/n}d^{-1/2}\right). \tag{23}$$

To bound $E_V(L_3)$, notice that

$$\begin{aligned}
E_V(L_3) &= \frac{1}{\sigma^2}E_V\left[h_{opt}(\hat\eta^2)\{1-h_{opt}(\hat\eta^2)\}\boldsymbol\beta^T\Sigma(X^TX)^{-1}X^T\boldsymbol\epsilon\right] \\
&\quad -\frac{1}{\sigma^2}E_V\left[\frac{\eta^2 d/(n-d-1)}{\{\eta^2+d/(n-d-1)\}^2}\boldsymbol\beta^T\Sigma(X^TX)^{-1}X^T\boldsymbol\epsilon\right] \\
&\le \left[E_V\left|h_{opt}(\hat\eta^2)\{1-h_{opt}(\hat\eta^2)\}-\frac{\eta^2 d/(n-d-1)}{\{\eta^2+d/(n-d-1)\}^2}\right|^2\right]^{1/2} \\
&\quad \cdot\left\{\frac{1}{\sigma^6}E_V|\boldsymbol\beta^T\Sigma(X^TX)^{-1}X^T\boldsymbol\epsilon|^2\right\}^{1/2} \\
&\le 2\left[E_V\left|\left\{h_{opt}(\hat\eta^2)-\frac{\eta^2}{\eta^2+d/(n-d-1)}\right\}\{1-h_{opt}(\hat\eta^2)\}\right|^2\right. \\
&\quad \left.+\frac{\eta^2}{\eta^2+d/(n-d-1)}E_V\left|h_{opt}(\hat\eta^2)-\frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^2\right]^{1/2} \\
&\quad \cdot\left\{\frac{1}{\sigma^4}E_V|\boldsymbol\beta^T\Sigma(X^TX)^{-1}X^T\boldsymbol\epsilon|^2\right\}^{1/2}
\end{aligned}$$

and, by Lemma B5,

$$\sup_{V\in\Theta^d(\eta^2)} E_V(L_3) = O\left\{\left(\frac{\eta^2 d/n}{\eta^2+d/n}\right)d^{-1/2}\right\}.$$

The theorem follows by combining this with (22)-(23). □

Proof of Theorem 3. Suppose that $V\in\Theta^d(\eta^2)$ and consider the following decomposition of the absolute difference between the predictive risk of the oracle and adaptive James-Stein estimators, notice that

$$\begin{aligned}
\left|R_V\{\hat{\boldsymbol\beta}_{js}(\eta^2)\}-R_V\{\hat{\boldsymbol\beta}_{js}(\hat\eta_+^2)\}\right| &= |E_V(J_1+J_2-2J_3)| \\
&\le |E_V(J_1)|+|E_V(J_2)|+2|E_V(J_3)|, \qquad (24)
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \frac{1}{\sigma^2}\left[\left\{\frac{\eta^2}{\eta^2+d/(n-d-1)}\right\}^2-\left\{\frac{\hat\eta_+^2}{\hat\eta_+^2+d/(n-d-1)}\right\}^2\right] \\
&\quad \cdot||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol\epsilon||^2, \\
J_2 &= \eta^2\left[\left\{\frac{d/(n-d-1)}{\eta^2+d/(n-d-1)}\right\}^2-\left\{\frac{d/(n-d-1)}{\hat\eta_+^2+d/(n-d-1)}\right\}^2\right],
\end{aligned}$$

$$J_3 = \frac{1}{\sigma^2}\frac{\hat{\eta}_+^2 d/(n-d-1)}{\{\hat{\eta}_+^2 + d/(n-d-1)\}^2}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}.$$

Similar to the proof of Theorem 2, we bound $|E_V(J_1)|$, $|E_V(J_2)|$, and $|E_V(J_3)|$ separately.

If $V \in \Theta^d(\eta^2)$, then

$$\begin{aligned}
|E_V(J_1)| &\leq E_V|J_1| \\
&= \frac{d}{n-d-1}E_V\left|\left[\frac{\eta^2-\hat{\eta}_+^2}{\{\eta^2+d/(n-d-1)\}\{\hat{\eta}_+^2+d/(n-d-1)\}}\right]\right. \\
&\qquad \cdot\left\{\frac{\eta^2}{\eta^2+d/(n-d-1)}+\frac{\hat{\eta}_+^2}{\hat{\eta}_+^2+d/(n-d-1)}\right\} \\
&\qquad \left.\cdot\frac{1}{\sigma^2}||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol{\epsilon}||^2\right|, \\
|E_V(J_2)| &\leq E_V|J_2| \\
&= \eta^2\left(\frac{d}{n-d-1}\right)^2 \\
&\qquad \cdot E_V\left|\left[\frac{\eta^2-\hat{\eta}_+^2}{\{\eta^2+d/(n-d-1)\}\{\hat{\eta}_+^2+d/(n-d-1)\}}\right]\right. \\
&\qquad \left.\cdot\left\{\frac{1}{\eta^2+d/(n-d-1)}+\frac{1}{\hat{\eta}_+^2+d/(n-d-1)}\right\}\right|.
\end{aligned}$$

Repeatedly applying the Cauchy-Schwarz inequality and Lemmas B2 and B4, it follows that

$$\sup_{V\in\Theta^d(\eta^2)}|E_V(J_1)|,\ \sup_{V\in\Theta^d(\eta^2)}|E_V(J_2)| = O\left\{\left(\frac{d/n}{\eta^2+d/n}\right)n^{-1/2}\right\}. \tag{25}$$

To bound $|E_V(J_3)|$, we use Stein's lemma (integration by parts). We have,

$$\begin{aligned}
E_V(J_3) &= \frac{1}{\sigma^2}E_V\left[\frac{\hat{\eta}_+^2 d/(n-d-1)}{\{\hat{\eta}_+^2+d/(n-d-1)\}^2}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}\right] \\
&= \frac{2}{n}E_V\left[\left\{\frac{d/(n-d-1)}{\hat{\eta}_+^2+d/(n-d-1)}\right\}^2\frac{1}{\hat{\sigma}^2}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\mathbf{y};\ \hat{\eta}_+^2>0\right] \\
&= \frac{2\eta^2}{n}E_V\left[\left\{\frac{d/(n-d-1)}{\hat{\eta}_+^2+d/(n-d-1)}\right\}^2\frac{\sigma^2}{\hat{\sigma}^2};\ \hat{\eta}_+^2>0\right] \\
&\quad +\frac{2}{n}E_V\left[\left\{\frac{d/(n-d-1)}{\hat{\eta}_+^2+d/(n-d-1)}\right\}^2\frac{1}{\hat{\sigma}^2}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon};\ \hat{\eta}_+^2>0\right].
\end{aligned}$$

Using Lemmas B2-B4, one easily checks that

$$\sup_{V\in\Theta^d(\eta^2)}\frac{2\eta^2}{n}E_V\left|\left\{\frac{d/(n-d-1)}{\hat{\eta}_+^2+d/(n-d-1)}\right\}^2\frac{\sigma^2}{\hat{\sigma}^2}\right| = O\left\{\left(\frac{d/n}{\eta^2+d/n}\right)n^{-1/2}\right\}$$

$$\sup_{V\in\Theta^d(\eta^2)}\frac{2}{n}E_V\left|\left\{\frac{d/(n-d-1)}{\hat{\eta}_+^2+d/(n-d-1)}\right\}^2\cdot\frac{1}{\hat{\sigma}^2}\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}\right| = O\left\{\left(\frac{d/n}{\eta^2+d/n}\right)n^{-1/2}\right\}.$$

We conclude that

$$\sup_{V\in\Theta^d(\eta^2)}|E_V(J_3)| = O\left\{\left(\frac{d/n}{\eta^2+d/n}\right)n^{-1/2}\right\}.$$

The theorem follows by combining this with (24)-(25). □

Proof of Proposition 3. Let $V\in\Theta^d(\eta^2)$ and suppose that $\Sigma^{-1}=USU^T$, where $U\in O(d)$ and $\mathrm{diag}(s_1,\ldots,s_d)$ is a diagonal matrix. Then

$$\begin{aligned}
R_V\{\hat{\boldsymbol{\beta}}_r(\lambda I)\} &= \sigma^{-2}E_V\left[\{\hat{\boldsymbol{\beta}}_r(\lambda I)-\boldsymbol{\beta}\}^T\Sigma\{\hat{\boldsymbol{\beta}}_r(\lambda I)-\boldsymbol{\beta}\}\right]\\
&= \sigma^{-2}E_V\left|\left|\Sigma^{1/2}\{(X^TX+n\lambda I)^{-1}X^TX-I\}\boldsymbol{\beta}\right|\right|^2\\
&\quad+\sigma^{-2}E_V\left|\left|\Sigma^{1/2}(X^TX+n\lambda I)^{-1}X^T\boldsymbol{\epsilon}\right|\right|^2\\
&= \sigma^{-2}E_I\left|\left|n\lambda(X^TX+n\lambda S)^{-1}SU^T\Sigma^{1/2}\boldsymbol{\beta}\right|\right|^2\\
&\quad+E_I\left[\mathrm{tr}\left\{(X^TX+n\lambda S)^{-2}X^TX\right\}\right].
\end{aligned}$$

Taking $s_1>0$ sufficiently large, $s_2,\ldots,s_d=1$, and $\boldsymbol{\beta}\in\mathbb{R}^d$ such that $\sigma^{-1}U^T\Sigma^{1/2}\boldsymbol{\beta}=(\eta,0,\ldots,0)\in\mathbb{R}^d$, one can ensure that

$$\sigma^{-2}E_I\left|\left|n\lambda(X^TX+n\lambda S)^{-1}SU^T\Sigma^{1/2}\boldsymbol{\beta}\right|\right|^2\geq\eta^2.$$

The proposition follows. □

Proof of Proposition 5. Let $s_1\geq\cdots\geq s_d\geq 0$ denote the eigenvalues of $n^{-1}X^TX$ and suppose that $V\in\Theta^d(\eta^2)$. Then, by Jensen's inequality and Proposition 4 (a),

$$\begin{aligned}
R_V\left[\hat{\boldsymbol{\beta}}_r\{d/(n\eta^2)\Sigma\}\right] &= E_I\left[\mathrm{tr}\left\{(X^TX+d/\eta^2 I)^{-1}\right\}\right]\\
&= E_I\left\{\frac{1}{n}\sum_{j=1}^d\frac{1}{s_j+d/(n\eta^2)}\right\}\\
&= \eta^2E_I\left\{\frac{1}{d}\sum_{j=1}^d\frac{s_j^{-1}}{n\eta^2/d+s_j^{-1}}\right\}\\
&\leq \eta^2\frac{E_I\mathrm{tr}(X^TX)^{-1}}{\eta^2+E_I\mathrm{tr}(X^TX)^{-1}}
\end{aligned}$$

$$\begin{aligned} &= \frac{\eta^2 d}{\eta^2(n-d-1)+d} \\ &= R_V(\hat{\boldsymbol{\beta}}^*_{js}), \end{aligned}$$

where the inequality is strict unless $\eta^2 = 0$. □

Appendix B

Lemma B1. *Suppose that $\kappa \geq 1$ is a fixed constant. Then*

$$\sup_{V \in \Theta^d(\eta^2)} \sigma^{2\kappa} E_V ||X\hat{\boldsymbol{\beta}}_{ols}||^{-2\kappa} = O\left\{ \left(\frac{1}{\eta^2 + d/n} \right)^{\kappa} n^{-\kappa} \right\}.$$

Proof. Conditional on X, the random variable $||X\hat{\boldsymbol{\beta}}_{ols}||^2/\sigma^2$ follows a noncentral χ^2 distribution with d degrees of freedom and noncentrality parameter $||X\boldsymbol{\beta}||^2/\sigma^2$. Thus, the distribution of $||X\hat{\boldsymbol{\beta}}_{ols}||^2/\sigma^2$ is the same as that of a central χ^2 random variable with $2N + d$ degrees of freedom, where $N|X \sim$ Poisson(ζ) and $\zeta = ||X\boldsymbol{\beta}||^2/(2\sigma^2)$. Since the κ-th inverse moment of a (central) χ^2 random with l degrees of freedom is $2^{-\kappa}\Gamma(l/2-\kappa)/\Gamma(l/2)$, provided $\kappa < l/2$, it follows that

$$\sigma^{2\kappa} E_V ||X\hat{\boldsymbol{\beta}}_{ols}||^{-2\kappa} = 2^{-\kappa} E_V \left\{ \frac{\Gamma(N + d/2 - \kappa)}{\Gamma(N + d/2)} \right\}$$

By Theorem 2.4 of (Ismail et al., 1986),

$$\frac{\Gamma(m-\kappa)}{\Gamma(m)} \leq (m-\kappa)^{-\kappa}$$

$m - \kappa > 0$. Thus,

$$\sigma^{2\kappa} E_V ||X\hat{\boldsymbol{\beta}}_{ols}||^{-2\kappa} \leq 2^{-\kappa} E_V (N + d/2 - \kappa)^{-\kappa} \leq (d - 2\kappa)^{-\kappa}$$

and

$$\begin{aligned} \sup_{V \in \Theta^d(\eta^2)} \sigma^{2\kappa} E_V ||X\hat{\boldsymbol{\beta}}_{ols}||^{-2\kappa} &\leq \sup_{V \in \Theta^d(\eta^2)} 2^{-\kappa} E_V (N + d/2 - \kappa)^{-\kappa} \\ &= O\left\{ \sup_{V \in \Theta^d(\eta^2)} \sigma^{2\kappa} E_V ||X\boldsymbol{\beta}||^{-2\kappa} \right\} \\ &= O(\eta^{-2\kappa} n^{-\kappa}). \end{aligned}$$

The lemma follows. □

Lemma B2. *Suppose that $\kappa \geq 1$ is a fixed constant and let $\hat{\eta}^2_+$ be as in (11). Suppose further that $0 < d/n \leq \rho_+ < 1$ for some fixed constant $\rho_+ \in \mathbb{R}$. Then*

$$\sup_{V \in \Theta^d(\eta^2)} E_V \left(\frac{1}{\hat{\eta}^2_+ + d/n} \right)^{\kappa} = O\left\{ \left(\frac{1}{\eta^2 + d/n} \right)^{\kappa} \right\}.$$

Proof. Notice that

$$\begin{aligned} E_V\left(\frac{1}{\hat{\eta}_+^2 + d/n}\right)^\kappa &\leq E_V\left\{\frac{\hat{\sigma}^2/\sigma^2}{||X\hat{\boldsymbol{\beta}}_{ols}||^2/(n\sigma^2)}\right\}^\kappa \\ &= E_V\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)^\kappa E_V\left\{\frac{1}{||X\hat{\boldsymbol{\beta}}_{ols}||^2/(n\sigma^2)}\right\}^\kappa. \end{aligned}$$

The result follows from Lemma B1. □

Lemma B3. *Suppose that* $0 < d/n \leq \rho_+ < 1$ *for some fixed constant* $\rho_+ \in \mathbb{R}$ *and let* $\kappa > 0$ *be fixed. Then*

$$\sup_{V \in \Theta^d(\eta^2)} P_V(\hat{\eta}_+^2 = 0) = O\left(\frac{d^{\kappa/2}}{\eta^{2\kappa} n^\kappa}\right).$$

Proof. Let $U = ||X\hat{\boldsymbol{\beta}}_{ols}||^2/\sigma^2$ and let $W = ||y - X\hat{\boldsymbol{\beta}}_{ols}||^2/\sigma^2 = (n-p)\hat{\sigma}^2/\sigma^2$. Then $W \sim \chi^2_{n-d}$ has a χ^2 distribution with $n-d$ degrees of freedom and, conditional on X, $U \sim \chi^2_{||X\boldsymbol{\beta}||^2/\sigma^2, d}$ has a noncentral χ^2 distribution with noncentrality parameter $||X\boldsymbol{\beta}||^2/\sigma^2$ and d degrees of freedom. Furthermore, U and W are independent and

$$\hat{\eta}_+^2 = \max\left\{\frac{d}{n}\left[\frac{U/d}{W/(n-d)} - 1\right], 0\right\}.$$

Thus, for $V \in \Theta^d(\eta^2)$,

$$\begin{aligned} P_V(\hat{\eta}_+^2 = 0) &= P_V\left(\frac{1}{d}U \leq \frac{1}{n-d}W\right) \\ &\leq E_V \exp\left(\frac{r}{n-d}W\right) E_V\left(-\frac{r}{d}U\right) \\ &= \left(\frac{1}{1 - \frac{2r}{n-d}}\right)^{(n-d)/2} \left(\frac{1}{1+\frac{2r}{d}}\right)^{d/2} \\ &\qquad \cdot E_V \exp\left(-\frac{r}{d+2r}||X\boldsymbol{\beta}||^2/\sigma^2\right) \\ &= \left(\frac{n-d}{n-d-2r}\right)^{(n-d)/2} \left(\frac{d}{d+2r}\right)^{d/2} \left\{\frac{d+2r}{d+2(\eta^2+1)r}\right\}^{n/2} \\ &\leq \exp\left\{\frac{2r^2 n}{(n-d-2r)(d+2r)}\right\} \left\{\frac{d+2r}{d+2(\eta^2+1)r}\right\}^{n/2}, \end{aligned}$$

provided $r < (n-d)/2$. Now, basic calculus implies that

$$\sup_{\eta^2 \geq 0} \eta^{2k}\left\{\frac{d+2r}{d+2(\eta^2+1)r}\right\}^{n/2} \leq e^{-k}\left\{\frac{(d+2r)k}{r(n-2k)}\right\}^k.$$

The lemma follows by taking $r = \alpha\sqrt{d}$ for $\alpha > 0$ sufficiently small. □

Lemma B4. *Suppose that $\rho_+, \kappa > 0$ are fixed constants and that $0 < d/n \leq \rho_+ < 1$. Then*

$$\sup_{V \in \Theta^d(\eta^2)} E_V |\hat{\eta}_+^2 - \eta^2|^\kappa = O\left\{ \frac{(d/n)^{\kappa/2} + \eta^\kappa + \eta^{2\kappa}}{n^{\kappa/2}} \right\}.$$

Proof. Using Lemma B3, we have

$$\begin{aligned} E_V |\hat{\eta}_+^2 - \eta^2|^\kappa &\leq E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\hat{\sigma}^2} - \left(\frac{d}{n} + \eta^2 \right) \right|^\kappa + \eta^{2\kappa} P_V(\hat{\eta}_+^2 = 0) \\ &= E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\hat{\sigma}^2} - \left(\frac{d}{n} + \eta^2 \right) \right|^\kappa + O\left(\frac{d^{\kappa/2}}{n^\kappa} \right). \end{aligned} \tag{26}$$

Since $(n-d)\hat{\sigma}^2 = ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ols}||^2$ and $||X\hat{\boldsymbol{\beta}}_{ols}||^2$ are independent,

$$\begin{aligned} E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\hat{\sigma}^2} - \left(\frac{d}{n} + \eta^2 \right) \right|^\kappa &\leq 2^\kappa E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\hat{\sigma}^2} - \left(\eta^2 + \frac{d}{n} \right) \frac{\sigma^2}{\hat{\sigma}^2} \right|^\kappa \\ &\quad + 2^\kappa \left(\eta^2 + \frac{d}{n} \right)^\kappa E_V \left| \frac{\sigma^2}{\hat{\sigma}^2} - 1 \right|^\kappa \\ &\leq 2^\kappa E_V \left(\frac{\sigma^2}{\hat{\sigma}^2} \right)^\kappa E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\sigma^2} - \left(\eta^2 + \frac{d}{n} \right) \right|^\kappa \\ &\quad + 2^\kappa \left(\eta^2 + \frac{d}{n} \right)^\kappa E_V \left| \frac{\sigma^2}{\hat{\sigma}^2} - 1 \right|^\kappa. \end{aligned}$$

As in the proof of Lemma B1, let $N \sim \text{Poisson}\{||X\boldsymbol{\beta}||^2/(2\sigma^2)\}$. Then, since $||X\hat{\boldsymbol{\beta}}_{ols}||^2/\sigma^2 \sim \chi^2_{2N+d}$,

$$\begin{aligned} E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\sigma^2} - \left(\eta^2 + \frac{d}{n} \right) \right|^\kappa &\leq 2^\kappa E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\sigma^2} - \frac{2N+d}{n} \right|^\kappa \\ &\quad + 2^\kappa E_V \left| \frac{2N}{n} - \eta^2 \right|^\kappa. \end{aligned}$$

Thus,

$$\sup_{V \in \Theta^d(\eta^2)} E_V \left| \frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\sigma^2} - \left(\eta^2 + \frac{d}{n} \right) \right|^\kappa = O\left\{ n^{-\kappa/2} \left(\eta^2 + \frac{d}{n} \right)^{\kappa/2} \right\}$$

Additionally, one can check that

$$\sup_{V \in \Theta^d(\eta^2)} E_V \left| \frac{\sigma^2}{\hat{\sigma}^2} - 1 \right|^\kappa = O(n^{-\kappa/2}).$$

It follows that

$$\sup_{V\in\Theta^d(\eta^2)} E_V\left|\frac{||X\hat{\boldsymbol{\beta}}_{ols}||^2}{n\hat{\sigma}^2}-\left(\frac{d}{n}+\eta^2\right)\right|^\kappa = O\left\{\frac{(d/n)^{\kappa/2}+\eta^\kappa+\eta^{2\kappa}}{n^{\kappa/2}}\right\}.$$

The lemma follows by combining this with (26). □

Lemma B5. *Suppose that $\rho_+,\kappa>0$ are fixed constants. Suppose further that $0<d/n\le\rho_+<1$ and let $h_{opt}(\hat{\eta}^2)$ be as in (7). Then*

$$\sup_{V\in\Theta^d(\eta^2)} E_V\left|h_{opt}(\hat{\eta}^2)-\frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^k = O\left\{\left(\frac{\eta}{\eta^2+d/n}\right)^k n^{-\kappa/2}\right\}.$$

Proof. Let

$$H=\left|h_{opt}(\hat{\eta}^2)-\frac{\eta^2}{\eta^2+d/(n-d-1)}\right|^\kappa.$$

Then

$$\begin{aligned} H &\le 2^\kappa\left|\frac{E_V\left(\boldsymbol{\beta}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}\middle|\hat{\eta}^2\right)-\eta^2\sigma^2}{E_V\left(||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^2\middle|\hat{\eta}^2\right)}\right|^\kappa \\ &\quad + 2^\kappa\left|\frac{\eta^2\left\{E_V\left(||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^2\middle|\hat{\eta}^2\right)-\eta^2\sigma^2-d\sigma^2/(n-d-1)\right\}}{\{\eta^2+d/(n-d-1)\}E_V\left(||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^2\middle|\hat{\eta}^2\right)}\right|^\kappa \end{aligned}$$

and, using the Cauchy-Schwarz and Jensen's inequalities,

$$\begin{aligned} E_V(H) &\le 2^\kappa\left\{\sigma^{4\kappa}E_V||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^{-4\kappa}\right\}^{1/2} \\ &\quad\cdot\left[\left\{E_V\left|\frac{\boldsymbol{\beta}^T\Sigma\hat{\boldsymbol{\beta}}_{ols}}{\sigma^2}-\eta^2\right|^{2\kappa}\right\}^{1/2}+\left\{\frac{\eta^2}{\eta^2+d/(n-d-1)}\right\}^\kappa\right. \\ &\quad\left.\cdot\left\{E_V\left|\frac{||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^2}{\sigma^2}-\eta^2-\frac{d}{n-d-1}\right|^{2\kappa}\right\}^{1/2}\right] \\ &\le 2^\kappa\left\{\sigma^{4\kappa}E_V||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^{-4\kappa}\right\}^{1/2}\left[(4^\kappa+1)\right. \\ &\quad\cdot\left\{E_V\left|\frac{\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}}{\sigma^2}\right|^{2\kappa}\right\}^{1/2}+\left\{\frac{2\eta^2}{\eta^2+d/(n-d-1)}\right\}^\kappa \\ &\quad\left.\cdot\left\{E_V\left|\frac{||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol{\epsilon}||^2}{\sigma^2}-\frac{d}{n-d-1}\right|^{2\kappa}\right\}^{1/2}\right]. \quad (27) \end{aligned}$$

Since $E_I||X^TX||^r = O(n^r)$ for fixed $r > 0$, where $||X^TX||$ denotes the operator norm of X^TX (see, for instance, Lemma C2 of (Dicker, 2013)), Lemma B1 implies that

$$\sup_{V\in\Theta^d(\eta^2)} \sigma^{4\kappa} E_V||\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{ols}||^{-4\kappa} = O\left\{\left(\frac{1}{\eta^2 + d/n}\right)^{2\kappa}\right\}. \tag{28}$$

Furthermore, since $E_I||(X^TX)^{-1}||^r = O(n^{-r})$ for fixed $r > 0$ (see, again, Lemma C2 of (Dicker, 2013)), it follows that

$$\sup_{V\in\Theta^d(\eta^2)} E_V\left|\frac{\boldsymbol{\beta}^T\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}}{\sigma^2}\right|^{2\kappa} = O(\eta^{2\kappa}n^{-\kappa}) \tag{29}$$

and

$$\sup_{V\in\Theta^d(\eta^2)} E_V\left|\frac{||\Sigma^{1/2}(X^TX)^{-1}X^T\boldsymbol{\epsilon}||^2}{\sigma^2} - \frac{d}{n-d-1}\right|^{2\kappa} = O(n^{-\kappa}). \tag{30}$$

Combining (27)-(30) yields

$$\begin{aligned}\sup_{V\in\Theta^d(\eta^2)} E_V(H) &= O\left[\left(\frac{1}{\eta^2 + d/n}\right)^{\kappa}\left\{\eta^{\kappa}n^{-\kappa/2} + \left(\frac{\eta^2}{\eta^2 + d/n}\right)^{\kappa} n^{-\kappa/2}\right\}\right] \\ &= O\left\{\left(\frac{\eta}{\eta^2 + d/n}\right)^{\kappa} n^{-\kappa/2}\right\},\end{aligned}$$

which proves the lemma. □

References

Baranchik, A. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Annals of Statistics* **1** 312–321. MR0348928

Beran, R. (1996). Stein estimation in high dimensions: A retrospective. In *Research developments in probability and statistics: Festschrift in honor of Madan L. Puri on the occasion of his 65th birthday*. VSP International Science Publishers. MR1462411

Brandwein, A. and Strawderman, W. (1990). Stein estimation: The spherically symmetric case. *Statistical Science* **5** 356–369. MR1080957

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78** 131–136. MR0696857

Brown, L. (1990). An ancillarity paradox which appears in multiple linear regression. *Annals of Statistics* **18** 471–493. MR1056325

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007a). Aggregation for gaussian regression. *Annals of Statistics* **35** 1674–1697. MR2351101

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007b). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194. MR2312149

Copas, J. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)* **45** 311–354. MR0737642

Dicker, L. (2013). Ridge regression and optimal dense estimation for high-dimensional linear models. Manuscript.

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

Huber, N. and Leeb, H. (2012). Shrinkage estimators for prediction out-of-sample: Conditional performance. ArXiv preprint arXiv:1209.0899. MR3031279

Ismail, M., Lorch, L. and Muldoon, M. (1986). Completely monotonic functions associated with the gamma function and its q-analogues. *Journal of Mathematical Analysis and Applications* **116** 1–9. MR0837337

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. MR0133191

Leeb, H. (2009). Conditional predictive inference post model selection. *Annals of Statistics* **37** 2838–2876. MR2541449

Marchand, E. (1993). Estimation of a multivariate mean with constraints on the norm. *Canadian Journal of Statistics* **21** 359–366. MR1254283

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc. MR0652932

Nussbaum, M. (1999). Minimax risk: Pinsker bound. In *Encyclopedia of Statistical Sciences*, vol. 3. Wiley, New York, 451–460.

Oman, S. (1984). A different empirical Bayes interpretation of ridge and Stein estimators. *Journal of the Royal Statistical Society: Series B (Methodological)* 544–557. MR0790637

Pinsker, M. (1980). Optimal filtration of functions from l2 in gaussian noise. *Problems of Information Transmission* **16** 52–68. MR0624591

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, vol. 1. MR0084922

Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. MR0120718

Sun, T. and Zhang, C. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

Takada, Y. (1979). A family of minimax estimators in some multiple regression problems. *Annals of Statistics* **7** 1144–1147. MR0536517

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

Tikhonov, A. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **39** 195–198. MR0009685