Bernoulli **22**(1), 2016, 1–37
DOI: 10.3150/14-BEJ609

Ridge regression and asymptotic minimax estimation over spheres of growing dimension

LEE H. DICKER

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

We study asymptotic minimax problems for estimating a d-dimensional regression parameter over spheres of growing dimension ($d \to \infty$). Assuming that the data follows a linear model with Gaussian predictors and errors, we show that ridge regression is asymptotically minimax and derive new closed form expressions for its asymptotic risk under squared-error loss. The asymptotic risk of ridge regression is closely related to the Stieltjes transform of the Marčenko–Pastur distribution and the spectral distribution of the predictors from the linear model. Adaptive ridge estimators are also proposed (which adapt to the unknown radius of the sphere) and connections with equivariant estimation are highlighted. Our results are mostly relevant for asymptotic settings where the number of observations, n, is proportional to the number of predictors, that is, $d/n \to \rho \in (0, \infty)$.

Keywords: adaptive estimation; equivariance; Marčenko–Pastur distribution; random matrix theory

1. Introduction

Consider a linear model where the observed data consists of outcomes $y_1, \ldots, y_n \in \mathbb{R}$ and d-dimensional predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ that are related via the equation

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

the d-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T \in \mathbb{R}^d$ is an unknown parameter and $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}$ are unobserved errors. To simplify notation, let $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$. Then (1) may be rewritten as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

In this paper, we study asymptotic minimax estimation of $\boldsymbol{\beta}$ over spheres of growing dimension (i.e., $d \to \infty$), under the assumption that the data $(\mathbf{y}, X)$ are jointly Gaussian. This is a variant of a problem considered by Goldenshluger and Tsybakov [31,32]; it is closely related to the fundamental work of Pinsker [43] and others, for example, Belitser and Levit [5], Beran [7], Golubev [33], on sharp asymptotic minimax estimation in the Gaussian sequence model. Taken together, the results in this paper provide a new example where sharp asymptotic minimax estimation is possible; an example that illustrates connections between linear models with many predictors and now classical results on the spectral distribution of large random matrices.

1.1. Statement of problem

Let I_k denote the $k \times k$ identity matrix. We assume throughout that

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} N(0, I_d) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim N(0, I_n) \tag{2}$$

are independent. More general models, where one might allow for positive definite $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ and arbitrary $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$, are discussed in Section 1.4.

Given an estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X)$ of $\boldsymbol{\beta}$, define the risk under squared-error loss

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\big(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\big), \tag{3}$$

where $\|\cdot\|$ denotes the ℓ^2-norm. The expectation in (3) is taken with respect to the joint distribution of $(X, \boldsymbol{\varepsilon})$ and the subscript $\boldsymbol{\beta}$ in $E_{\boldsymbol{\beta}}$ indicates that $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (for expectations that do not involve $\mathbf{y}$, we will often omit the subscript). We emphasize that the expectation in (3) is taken over the predictors X as well as the errors $\boldsymbol{\varepsilon}$; in other words, rather than conditioning on X, (3) is the *unconditional* risk under squared-error loss.

Let $S^{d-1}(\tau) = \{\boldsymbol{\beta} \in \mathbb{R}^d; \|\boldsymbol{\beta}\| = \tau\}$ be the sphere of radius $\tau \geq 0$ in $\mathbb{R}^d$ centered at the origin. The minimax risk for estimating $\boldsymbol{\beta}$ over $S^{d-1}(\tau)$ is given by

$$r(\tau) = r_{d,n}(\tau) = \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}), \tag{4}$$

where the infimum in (4) is taken over all measurable estimators $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X)$.

The minimax problem determined by (4) is the main focus of this paper. Our analysis entails (i) identifying and analyzing specific estimators $\hat{\boldsymbol{\beta}}$ such that $\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \approx r(\tau)$, and (ii) obtaining accurate closed-form approximations for $r(\tau)$, while focusing on settings where d is large.

1.2. Overview of results

To better orient the reader, we give a brief section-by-section overview of the paper. We conclude this section with an additional comment on the nature of the asymptotic results derived herein.

Section 2: *Ridge regression*. Ridge regression (Hoerl and Kennard [35] and Tihonov [50]) is a widely studied regularized estimation method whose use has been advocated in various settings where d is large or X is ill-conditioned. Our analysis in Section 2 yields a simple formula for the optimal ridge regularization parameter and a new closed-form expression for the associated ridge estimator's asymptotic risk. More specifically, we show that if $d/n \to \rho \in (0, \infty)$, then the asymptotic risk of the ridge estimator is closely related to the Stieltjes transform of the Marčenko–Pastur distribution (Marčenko and Pastur [39]), which plays a prominent role in random matrix theory, for example, Bai *et al.* [2], El Karoui [28], Silverstein [46]. Settings where $d/n \to 0$ and $d/n \to \infty$ are also considered. Our results for ridge regression immediately provide an upper bound on $r(\tau)$, in the usual way: It is clear from (4) that

$r(\tau) \leq \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for all estimators $\hat{\boldsymbol{\beta}}$; taking $\hat{\boldsymbol{\beta}}$ to be the specified ridge estimator gives the desired upper bound.

Section 3: *An equivalent Bayes problem*. An equivariance argument implies that $r(\tau)$ is equal to the Bayes risk for estimating $\boldsymbol{\beta}$ under the prior distribution $\boldsymbol{\beta} \sim \pi_{S^{d-1}(\tau)}$, where $\pi_{S^{d-1}(\tau)}$ denotes the uniform distribution on $S^{d-1}(\tau)$ (this is an application of well-known results on equivariance, e.g., Chapter 6 of Berger [8], and is essentially an illustration of the Hunt–Stein theorem (Bondar and Milnes [11])). Additionally, we argue that when d is large, the Bayes risk for estimating $\boldsymbol{\beta}$ under the prior distribution $\boldsymbol{\beta} \sim \pi_{S^{d-1}(\tau)}$ is close to the Bayes risk for estimating $\boldsymbol{\beta}$ under a normal prior distribution, which coincides with the risk of ridge regression. We conclude that the risk of ridge regression is asymptotically equivalent to $r(\tau)$ and that ridge regression is asymptotically optimal for estimation over $S^{d-1}(\tau)$.

Section 4: *An adaptive ridge estimator*. The ridge regression estimator $\hat{\boldsymbol{\beta}}_r(\tau)$ that is asymptotically optimal over $S^{d-1}(\tau)$ depends on the radius $\tau = \|\boldsymbol{\beta}\|$, which is typically unknown. Replacing τ with an estimate, we obtain an adaptive ridge estimator that does not depend on τ, but is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_r(\tau)$. It follows that the adaptive ridge estimator is adaptive asymptotic minimax over spheres $S^{d-1}(\tau)$, provided $\tau^2 \gg n^{-1/2}$. Additionally, we show that the adaptive ridge estimator is asymptotically optimal among the class of all estimators for $\boldsymbol{\beta}$ that are equivariant with respect to orthogonal transformations of the predictors, as $d \to \infty$.

Proofs may be found in the Appendices.

Note on asymptotics. Throughout the paper, our asymptotic analysis is focused on settings where $d \to \infty$. We typically assume that $n \to \infty$ along with d and that $d/n \to \rho \in [0, \infty]$. It will become apparent below that most of the "action" occurs when $0 < \rho < \infty$. Indeed, one of the implications of our results is that if $0 < \rho < \infty$, then the minimax risk $r(\tau)$ is influenced by the spectral distribution of the empirical covariance matrix $n^{-1}X^T X$. On the other hand, if $\rho = 0$, then the behavior of $r(\tau)$ is more standard. If $\rho = \infty$, then we will show that it is impossible to out-perform the trivial estimator $\hat{\boldsymbol{\beta}}_{\text{null}} = 0$ for estimation over $S^{d-1}(\tau)$; note the contrast with sparse estimation problems, where $\boldsymbol{\beta}$ is assumed to be sparse and it may be possible to dramatically out-perform $\hat{\boldsymbol{\beta}}_{\text{null}}$ when $d/n \to \infty$, for example, Bickel *et al.* [10], Bunea *et al.* [18], Candes and Tao [19], Raskutti *et al.* [44], Ye and Zhang [52], Zhang [54].

1.3. Relationship to existing work

The minimax problem (4) is closely related to problems considered by Goldenshluger and Tsybakov [31,32], who studied minimax prediction problems over ℓ^2-ellipsoids

$$\left\{ \boldsymbol{\beta} \in \ell^2; \sum_{k=1}^{\infty} a_k^2 \beta_k^2 \leq L^2 \right\}; \qquad L > 0, a = \{a_k\},$$

in an infinite-dimensional linear model with independent (but not necessarily Gaussian) predictors. Goldenshluger and Tsykbakov's results apply to classes of ellipsoids with various constraints on $a = \{a_k\}$ and L. Taking $L = \tau$, $a_1 = \cdots = a_d = 1$, and $a_{d+1} = a_{d+2} = \cdots = \infty$ (and following the convention that $0 \times \infty = 0$), the results in Goldenshluger and Tsybakov

[32] may be applied to obtain asymptotics for the minimax risk over the d-dimensional ball $B_d(\tau) = \{\boldsymbol{\beta} \in \mathbb{R}^d; \|\boldsymbol{\beta}\| \leq \tau\}$,

$$\bar{r}(\tau) = \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in B_d(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}); \tag{5}$$

results in Goldenshluger and Tsybakov [31] yield adaptive estimators that are asymptotically minimax over classes of balls $B_d(\tau)$. In Section 3.2, we show that $r(\tau) \approx \bar{r}(\tau)$, when d is large (see (12) below). Thus, Goldenshluger and Tsybakov's results are clearly related to the results presented here. However, as applied to balls $B_d(\tau)$, their results typically require that $d/n \to 0$ (for instance, Theorem 1 of Goldenshluger and Tsybakov [32] requires that $d = o\{\sqrt{n/\log(n)}\}$ and Assumption 3 of Goldenshluger and Tsybakov [31] requires $d = O\{\sqrt{n}/\log(n)\}$). By contrast, the results in this paper apply in settings where $d/n \to \rho \in [0, \infty]$, with the bulk of our work focusing on $0 < \rho < \infty$.

The analysis in this paper focuses on estimation over the sphere $S^{d-1}(\tau)$, rather than the ball $B_d(\tau)$; that is, we focus on the minimax problem (4), as opposed to (5). The ball $B_d(\tau)$ and other star-shaped parameter spaces (e.g., ellipsoids or ℓ^p-balls) have been more frequently studied in the literature on asymptotic minimax problems over restricted parameter spaces (Donoho and Johnstone [26], Golubev [34], Nussbaum [42]). Evidently, the problems (4) and (5) are closely related. However, analysis of (4) appears to be somewhat more complex; in particular, obtaining lower bounds on $r(\tau)$ seems more challenging. To justify our emphasis on the sphere $S^{d-1}(\tau)$, in Section 3.2, we show that asymptotics for $\bar{r}(\tau)$ follow easily from asymptotics for $r(\tau)$. Additionally, by studying estimation over the sphere, we are able to draw deeper connections with equivariance than seem to be available if one focuses on the ball (e.g., Proposition 7 below). A similar approach has been considered by Marchand [40] and Beran [7] in their analysis of the finite-dimensional Gaussian sequence mode. In fact, one of the key technical results in this paper (Theorem 2) is essentially a multivariate extension of Theorem 3.1 in Marchand [40]. While we believe that the additional insights provided by studying minimax problems over the sphere justify the added complexity, we also note that more standard approaches to obtaining lower bounds on the minimax risk over balls (see, e.g., Nussbaum [42] or Chapter 3 of Tsybakov [51]) may be applied to obtain lower bounds for $\bar{r}(\tau)$ directly.

Finally in this section, we mention some of the existing work on random matrix theory that is especially relevant for our analysis of ridge regression in Section 2. Theorem 1 in Section 2.2 relies heavily on now classical results that describe the asymptotic behavior of the empirical distribution of the eigenvalues of $n^{-1}X^TX$ in high dimensions (Bai [1], Bai *et al.* [2], Marčenko and Pastur [39]). Additionally, we point out that while other authors have alluded to the relevance of random matrix theory for ridge regression (El Karoui and Kösters [29]), the results presented here on ridge regression's asymptotic risk seem to provide a greater level of detail than available elsewhere, in the specified setting.

1.4. Distributional assumptions

The linear model (1) with distributional assumptions (2) is highly specialized. However, similar models have been studied previously. Stein [48], Baranchik [3], Breiman and Freedman [13], Brown [15] and Leeb [37] studied estimation problems for linear models with jointly Gaussian

data, but, for the most part, these authors do not require $\mathrm{Cov}(\mathbf{x}_i) = I_d$. Moreover, as discussed in Section 1.3, the infinite-dimensional linear model considered by Goldenshluger and Tsybakov [31,32] is similar to the model studied in this paper. For our purposes, one of the more significant consequences of the normality assumption (2) is that the distributions of X and $\boldsymbol{\varepsilon}$ are invariant under orthogonal transformations. This leads to substantial simplifications in many of the ensuing calculations. Results in El Karoui and Kösters [29] suggest that a general approach to relaxing some of the distributional assumptions made in this paper may be feasible, but this is not pursued further here.

We point out that the assumption $E(\mathbf{x}_i) = 0$, which is implicit in (2), is not particularly limiting: If $E(\mathbf{x}_i) \neq 0$, then we can reduce to the mean 0 case by centering and de-correlating the data. The normality assumption (2) also requires $\mathrm{Var}(\varepsilon_i) = 1$. If $\mathrm{Var}(\varepsilon_i) = \sigma^2 \neq 1$ and σ^2 is known, then this can be reduced to the case where $\mathrm{Var}(\varepsilon_i) = 1$ by transforming the data $(\mathbf{y}, X) \mapsto (\mathbf{y}/\sigma, X)$; the corresponding transformation for the parameters $\boldsymbol{\beta}$, σ^2 is given by $(\boldsymbol{\beta}, \sigma^2) \mapsto (\boldsymbol{\beta}/\sigma, 1)$ and the risk function should be scaled by σ^2, as well (ultimately in this scenario, most of the results in this paper remain valid except that the signal-to-noise ratio $\|\boldsymbol{\beta}\|^2/\sigma^2$ replaces the signal strength $\|\boldsymbol{\beta}\|^2$). If σ^2 is unknown and $d/n \to \rho < 1$, then σ^2 may be effectively estimated by $\hat{\sigma}^2 = (n-d)^{-1}\|y - X\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\|^2$, where $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = (X^T X)^{-1} X^T \mathbf{y}$ is the ordinary least squares (OLS) estimator; one can subsequently reduce to the case where $\mathrm{Var}(\varepsilon_i) = 1$. (Throughout, if the square matrix A is not invertible, then we take A^{-1} to be its Moore–Penrose pseudoinverse; typically, the matrices we seek to invert will be invertible with probability 1.) Recent work suggests that σ^2 may also be effectively estimated when $d > n$. Fan *et al.* [30] and Sun and Zhang [49] propose methods for estimating σ^2 when $d > n$ and $\boldsymbol{\beta}$ is sparse (see also related work by Belloni *et al.* [6] and Dalalyan and Chen [20] on estimating $\boldsymbol{\beta}$ in high dimensions when σ^2 is unknown); Dicker [25] considers estimating σ^2 when $d > n$ and $\boldsymbol{\beta}$ is not sparse.

Under the Gaussian assumption (2), the predictors $\mathbf{x}_i$ are uncorrelated at the population level, that is, $\mathrm{Cov}(\mathbf{x}_i) = I_d$. The results in this paper are easily adapted to settings where $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is a known positive definite matrix by transforming the data $(\mathbf{y}, X) \mapsto (\mathbf{y}, X\Sigma^{-1/2})$, and making corresponding transformations of the parameters and risk function. If $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is unknown, but $\hat{\Sigma}$ is an operator norm consistent estimator, then it is straigthforward to check that most of our asymptotic results remain valid, mutatis mutandis, for the transformed data $(\mathbf{y}, X\hat{\Sigma}^{-1/2})$. On the other hand, in high-dimensional settings where $d/n \to \rho > 0$, an operator norm consistent estimator for Σ may not exist. In Dicker [24], the author considers a prediction problem closely related to the estimation problem considered in this paper, with unknown $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$; the author identifies an asymptotically optimal equivariant estimator and derives expressions for the estimator's asymptotic risk (Theorems 2–3 and Corollary 1 of Dicker [24]). One interpretation of the results in Dicker [24] is that they quantify the loss in efficiency of equivariant estimators when $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is unknown, as compared to the results presented here for the case where $\mathrm{Cov}(\mathbf{x}_i) = I_d$ is known.

2. Ridge regression

Define the ridge regression estimator

$$\hat{\boldsymbol{\beta}}_r(t) = \left(X^T X + d/t^2 I_d\right)^{-1} X^T \mathbf{y}, \qquad t \in [0, \infty].$$

The parameter t is referred to as the "regularization" or "ridge" parameter and is subject to further specification. By convention, we take $\hat{\boldsymbol{\beta}}_r(0) = 0$ and $\hat{\boldsymbol{\beta}}_r(\infty) = \hat{\boldsymbol{\beta}}_{\text{ols}} = (X^T X)^{-1} X^T \mathbf{y}$ to be the OLS estimator.

2.1. The oracle ridge estimator

Our first result identifies the optimal ridge parameter t and yields an oracle ridge estimator with minimal risk. A simplified expression for the oracle ridge estimator's risk is also provided.

Proposition 1. *Suppose that* $\boldsymbol{\beta} \in S^{d-1}(\tau)$. *Then*

$$R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} = \inf_{t \in [0,\infty]} R\big\{\hat{\boldsymbol{\beta}}_r(t), \boldsymbol{\beta}\big\} = E\Big[\text{tr}\Big\{\big(X^T X + d/\tau^2 I_d\big)^{-1}\Big\}\Big]. \tag{6}$$

Corollary 1. *Suppose that* $\tau \geq 0$. *Then*

$$r(\tau) \leq \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} = E\Big[\text{tr}\Big\{\big(X^T X + d/\tau^2 I_d\big)^{-1}\Big\}\Big].$$

Proposition 1 is proved in Appendix A and it implies that the optimal ridge parameter is given by the signal strength $\tau = \|\boldsymbol{\beta}\|$. Notice that the risk of $\hat{\boldsymbol{\beta}}_r(\tau)$ is constant over the sphere $\boldsymbol{\beta} \in S^{d-1}(\tau)$. Corollary 1, which gives an upper bound on $r(\tau)$, follows immediately from Proposition 1 and the definition of $r(\tau)$.

In practice, the signal strength $\tau = \|\boldsymbol{\beta}\|$ is typically unknown. Thus, with $\boldsymbol{\beta} \in S^{d-1}(\tau)$, $\hat{\boldsymbol{\beta}}_r(\tau)$ may be viewed as an oracle estimator. In cases where the signal strength is not prespecified, Proposition 1 implies that $\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|)$ is the oracle estimator with minimal risk among ridge estimators. We will refer to both $\hat{\boldsymbol{\beta}}_r(\tau)$ and $\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|)$ as the oracle ridge estimator, according to whether or not $\boldsymbol{\beta} \in S^{d-1}(\tau)$ has been specified in advance. In Section 4, we discuss adaptive ridge estimators that utilize an estimate of the signal strength.

Expressions similar to those in Proposition 1 for the optimal ridge parameter and the risk ridge estimators have appeared previously in the literature (see, e.g., the review article by Draper and Van Nostrand [27]). However, other existing results on the risk of ridge estimators tend to either (i) be significantly more complex than Proposition 1 or (ii) pertain to the Bayes risk of ridge regression, assuming that $\boldsymbol{\beta}$ follows a normal prior distribution. Proposition 1 is a simple, yet conclusive result for the optimal ridge parameter with respect to the frequentist risk $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. Its simplicity follows largely from the symmetry in our formulation of the problem; in particular, we are focusing on unconditional risk and the distribution of X is orthogonally invariant.

2.2. Asymptotic risk

It appears that the risk formula (6) cannot be further simplified with ease. However, results from random matrix theory yield a closed-form expression for the asymptotic risk. For $\rho \in (0, \infty)$, the

Marčenko–Pastur density f_ρ is defined by

$$f_\rho(z) = \max\{(1-\rho^{-1}), 0\}\delta_0(z) + \frac{1}{2\pi\rho z}\sqrt{(b-z)(z-a)}\mathbf{1}_{(a,b)}(z),$$

where $a = (1-\sqrt{\rho})^2$, $b = (1+\sqrt{\rho})^2$, $\delta_0(\cdot)$ is the Dirac delta, and $\mathbf{1}_{(a,b)}(\cdot)$ is the indicator function of the open interval (a, b). The density f_ρ determines the Marčenko–Pastur distribution, which is the limiting distribution of the eigenvalues of $n^{-1}X^TX$, if $n \to \infty$ and $d/n \to \rho \in (0, \infty)$ (Marčenko and Pastur [39]); it also determines the corresponding cumulative distribution function, $F_\rho(t) = \int_{-\infty}^t f_\rho(z)\,\mathrm{d}z$. The Stieltjes transform of the Marčenko–Pastur distribution is defined by

$$\begin{aligned} m_\rho(s) &= \int \frac{1}{z-s} f_\rho(z)\,\mathrm{d}z = \int \frac{1}{z-s}\,\mathrm{d}F_\rho(z) \\ &= -\frac{1}{2\rho s}\{s+\rho-1+\sqrt{(s+\rho-1)^2-4\rho s}\}, \qquad s<0. \end{aligned} \tag{7}$$

The main result of this section implies that if $\boldsymbol{\beta} \in S^{d-1}(\tau)$, then the risk of the oracle ridge estimator may be approximated by $(d/n)m_{d/n}\{-d/(n\tau^2)\}$.

Theorem 1. *Suppose that* $0 < \rho_- \le d/n \le \rho_+ < \infty$ *for some fixed constants* $\rho_-, \rho_+ \in \mathbb{R}$.

(a) *If* $0 < \rho_- < \rho_+ < 1$ *or* $1 < \rho_- < \rho_+ < \infty$ *and* $|n-d| > 5$, *then*

$$\sup_{\boldsymbol{\beta}\in S^{d-1}(\tau)}\left|R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - \frac{d}{n}m_{d/n}\left(-\frac{d}{n\tau^2}\right)\right| = \mathrm{O}\left(\frac{\tau^2}{1+\tau^2}n^{-1/2}\right).$$

(b) *If* $0 < \rho_- < 1 < \rho_+ < \infty$, *then*

$$\sup_{\boldsymbol{\beta}\in S^{d-1}(\tau)}\left|R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - \frac{d}{n}m_{d/n}\left(-\frac{d}{n\tau^2}\right)\right| = \mathrm{O}(\tau^2 n^{-1/8}).$$

Theorem 1 is proved in Appendix A. Since $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}$ is constant over $\boldsymbol{\beta} \in S^{d-1}(\tau)$, the supremums in parts (a) and (b) of Theorem 1 are somewhat superfluous; however, they serve to emphasize that the upper bounds do not depend on any particular value of $\boldsymbol{\beta} \in S^{d-1}(\tau)$.

Let $0 \le s_d \le s_{d-1} \le \cdots \le s_1$ denote the ordered eigenvalues of $n^{-1}X^TX$ and define the empirical cumulative distribution function $\mathbb{F}_{n,d}(s) = d^{-1}\sum_{j=1}^d \mathbf{1}_{(-\infty,s_j]}(s)$. There are two keys to the proof of Theorem 1. The first is the observation that if $\boldsymbol{\beta} \in S^{d-1}(\tau)$, then, by Proposition 1,

$$\frac{n}{d}R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} = \frac{1}{d}E\left[\mathrm{tr}\left\{\left(\frac{1}{n}X^TX + \frac{d}{n\tau^2}I_d\right)^{-1}\right\}\right] = E\left\{\int \frac{1}{s+d/(n\tau^2)}\,\mathrm{d}\mathbb{F}_{n,d}(s)\right\};$$

in other words, the risk of the oracle ridge estimator is the expected value of the Stieltjes transform of $\mathbb{F}_{n,d}$. The second key is Theorem 1.1 of Bai *et al.* [2], which states that under the condi-

tions of Theorem 1,

$$\sup_{s\in\mathbb{R}}\big|E\big\{\mathbb{F}_{n,d}(s)\big\}-F_{d/n}(s)\big|=\begin{cases}\mathrm{O}\big(n^{-1/2}\big) & \text{if } 0<\rho_-<\rho_+<1 \text{ or } 1<\rho_-<\rho_+<\infty,\\ \mathrm{O}\big(n^{-1/8}\big) & \text{if } 0<\rho_-<1<\rho_+<\infty.\end{cases} \tag{8}$$

The different rates in (8) depending on whether or not $\rho_- < 1 < \rho_+$ helps to explain why these situations are considered separately in Theorem 1 above; more fundamentally, the major difference between the two cases is that if $d/n \to 1$ (corresponding to the setting where $\rho_- < 1 < \rho_+$), then 0 is contained in the support of the continuous part of the Marčenko–Pastur distribution, which complicates the analysis.

The asymptotic risk of the oracle ridge estimator, when $d/n \to \rho \in (0,\infty)$, is given explicitly in the following corollary, which follows immediately from Theorem 1.

Corollary 2. *For $\rho \in (0,\infty)$ and $\tau \in [0,\infty)$ define the asymptotic risk of the oracle ridge estimator*

$$R_r(\tau,\rho)=\frac{1}{2\rho}\big[\tau^2(\rho-1)-\rho+\sqrt{\big\{\tau^2(\rho-1)-\rho\big\}^2+4\rho^2\tau^2}\big].$$

(a) *If $\rho \in (0,\infty)\setminus\{1\}$, then*

$$\lim_{d/n\to\rho}\sup_{\boldsymbol{\beta}\in\mathbb{R}^d}\big|R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big),\boldsymbol{\beta}\big\}-R_r\big(\|\boldsymbol{\beta}\|,d/n\big)\big|=0.$$

(b) *If $0 \le T < \infty$ is a fixed real number, then*

$$\lim_{d/n\to 1}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{R}^d;\\ 0\le\|\boldsymbol{\beta}\|\le T}}\big|R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big),\boldsymbol{\beta}\big\}-R_r\big(\|\boldsymbol{\beta}\|,d/n\big)\big|=0.$$

In Corollary 2 and throughout the paper, the notation $\lim_{d/n\to\rho}$ indicates the limit as $n\to\infty$ and $d/n\to\rho$. Corollary 2 implies that if $d/n \to \rho \in (0,\infty)\setminus\{1\}$, then the risk of the oracle ridge estimator $\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|)$ converges to the asymptotic risk $R_r(\|\boldsymbol{\beta}\|,d/n)$ uniformly over all $\boldsymbol{\beta}\in\mathbb{R}^d$; if $d/n\to 1$, then the convergence is uniform over compact sets.

It is clear from Theorem 1 and Corollary 2 that if $d/n\to\rho\in(0,\infty)$, then the spectral distribution of $n^{-1}X^TX$ plays a prominent role in determining the risk of the oracle ridge estimator via the Marčenko–Pastur law; if $d/n\to 0$ or $d/n\to\infty$, then its role subsides, as illustrated by the following proposition.

Proposition 2.

(a) $[d/n\to 0]$ *For $\rho,\tau\in[0,\infty)$ define*

$$R_r^0(\tau,\rho)=\frac{\rho\tau^2}{\rho+\tau^2}. \tag{9}$$

Then

$$\lim_{d/n\to 0}\sup_{\boldsymbol{\beta}\in\mathbb{R}^d}\left|\frac{R\{\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|),\boldsymbol{\beta}\}}{R_r^0(\|\boldsymbol{\beta}\|,d/n)}-1\right|=0.$$

(b) $[d/n\to\infty]$ *Let* $0\le T<\infty$ *be a fixed real number. Then*

$$\lim_{d/n\to\infty}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{R}^d;\\ 0\le\|\boldsymbol{\beta}\|\le T}}\left|R\{\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|),\boldsymbol{\beta}\}-\|\boldsymbol{\beta}\|^2\right|=0.$$

Proposition 2 is proved in Appendix A. It gives the asymptotic risk of the oracle ridge estimator in settings where $d/n\to 0$ and ∞. Expressions like (9) are common in the analysis of linear estimators for the Gaussian sequence model (Pinsker [43]). Thus, if $d/n\to 0$, then features of $R\{\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|),\boldsymbol{\beta}\}$ deriving from the random predictors X are less apparent.

Now consider the null estimator $\hat{\boldsymbol{\beta}}_{\text{null}}=0$ and notice that $R(\hat{\boldsymbol{\beta}}_{\text{null}},\boldsymbol{\beta})=\|\boldsymbol{\beta}\|^2$. Proposition 2(b) implies that if $d/n\to\infty$, then the oracle ridge estimator is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\text{null}}$. In Section 3, we argue that if $d/n\to\infty$, then $\hat{\boldsymbol{\beta}}_{\text{null}}$ is in fact asymptotically minimax for the problem (4). In other words, non-trivial estimation is impossible in (4) when $d/n\to\infty$.

Combined with Theorem 1, Proposition 2 implies that the asymptotic risk of the oracle ridge estimator $R_r(\tau,\rho)$ extends continuously to $\rho=0$ and $\rho=\infty$. For $\tau\ge 0$, we define $R_r(\tau,0)=0$ and $R_r(\tau,\infty)=\tau^2$.

3. An equivalent Bayes problem

In this section, we use an equivariance argument to reduce the minimax problem (4) to an equivalent Bayes problem. We then show that ridge regression solves the Bayes problem, asymptotically.

3.1. The uniform measure on $S^{d-1}(\tau)$ and equivariance

Let $\pi_{S^{d-1}(\tau)}$ denote the uniform measure on $S^{d-1}(\tau)$. Define the Bayes risk

$$r_B(\tau)=\inf_{\hat{\boldsymbol{\beta}}}\int_{S^{d-1}(\tau)}R(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})\,\mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta})=\inf_{\hat{\boldsymbol{\beta}}}E_{\pi_{S^{d-1}(\tau)}}\big(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2\big),\tag{10}$$

where the expectation $E_{\pi_{S^{d-1}(\tau)}}$ is taken with respect to the joint distribution of $(X,\boldsymbol{\varepsilon},\boldsymbol{\beta})$, with $\boldsymbol{\beta}\sim\pi_{S^{d-1}(\tau)}$ independent of $(X,\boldsymbol{\varepsilon})$. The Bayes estimator

$$\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}=E_{\pi_{S^{d-1}(\tau)}}(\boldsymbol{\beta}|\mathbf{y},X)$$

satisfies

$$r_B(\tau)=E_{\pi_{S^{d-1}(\tau)}}\big\{\|\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}-\boldsymbol{\beta}\|^2\big\}.\tag{11}$$

Let $\mathcal{O}(d)$ denote the group of $d \times d$ orthogonal matrices. As with $\boldsymbol{\varepsilon}$ and X, the distribution $\pi_{S^{d-1}(\tau)}$ is invariant under orthogonal transformations; that is, if $U \in \mathcal{O}(d)$ and $\boldsymbol{\beta} \sim \pi_{S^{d-1}(\tau)}$, then $U\boldsymbol{\beta} \sim \pi_{S^{d-1}(\tau)}$. A corresponding feature of the estimator $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$ is that it is *equivariant* with respect to orthogonal transformations.

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

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

for all $d \times d$ *orthogonal matrices* $U \in \mathcal{O}(d)$.

Let

$$\mathscr{E} = \mathscr{E}_{d,n} = \{\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}} \text{ is an orthogonally equivariant estimator for } \boldsymbol{\beta}\}.$$

Then one easily checks that $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)} \in \mathscr{E}$. Additionally, notice that $\hat{\boldsymbol{\beta}}_r(\tau) \in \mathscr{E}$ is orthogonally equivariant. The following proposition is proved in Appendix A.

Proposition 3. *Suppose that* $\tau \geq 0$ *and that* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in S^{d-1}(\tau)$.

(a) *If* $\hat{\boldsymbol{\beta}}$ *is an orthogonally equivariant estimator, then the risk of* $\hat{\boldsymbol{\beta}}$ *is constant over* $S^{d-1}(\tau)$; *that is,* $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1) = R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$.

(b)

$$r(\tau) = \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}_1\} = E_{\pi_{S^{d-1}(\tau)}}\left\{\|\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)} - \boldsymbol{\beta}\|^2\right\} = r_B(\tau).$$

Proposition 3(a) implies that all orthogonally equivariant estimators have constant risk over spheres $S^{d-1}(\tau)$; we first noted that ridge regression possesses this property in a remark following Proposition 1. Proposition 3(b) implies that the Bayes problem (10) and the minimax problem (4) are equivalent. Proposition 3(b) also implies that the estimator $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$ is minimax over $S^{d-1}(\tau)$. While this, in a sense, "solves" the main problem of interest (4), there are several caveats. For instance, the estimator $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$ is an oracle estimator (it depends on τ) and is difficult to compute, even if τ is known. Furthermore, Proposition 3 provides no information about the magnitude of $r(\tau)$. In the next section, we show that when d is large, $r(\tau) = R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} \approx R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} \approx R_r(\tau, \rho)$ for $\boldsymbol{\beta} \in S^{d-1}(\tau)$. In addition to providing quantitative information about $r(\tau)$, this result suggests that ridge regression may be an appealing alternative to $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$, especially when combined with results on adaptive ridge estimators in Section 4.

3.2. Ridge regression and asymptotic optimality

Recall that the minimax estimator $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$ is the posterior mean of $\boldsymbol{\beta}$, under the assumption that $\boldsymbol{\beta} \sim \pi_{S^{d-1}(\tau)}$ is uniformly distributed on the sphere $S^{d-1}(\tau)$. On the other hand, the oracle

ridge estimator $\hat{\boldsymbol{\beta}}_r(\tau) = E_{N(0,\tau^2/dI_d)}(\boldsymbol{\beta}|\mathbf{y}, X)$ may be interpreted as the posterior mean of $\boldsymbol{\beta}$ under the assumption that $\boldsymbol{\beta} \sim N(0, \tau^2/dI_d)$ is normally distributed and independent of $(X, \boldsymbol{\varepsilon})$. If d is large, then the normal distribution $N(0, \tau^2/dI_d)$ is "close" to the uniform distribution on $S^{d-1}(\tau)$ (there is an enormous body of literature that makes this idea more precise – Diaconis and Freedman [23] attribute early work to Borel [12] and Lévy [38]). Thus, it is reasonable to expect that if d is large and $\boldsymbol{\beta} \in S^{d-1}(\tau)$, then $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)} \approx \hat{\boldsymbol{\beta}}_r(\tau)$ and that the two estimators have similar risk properties. This is the content of the main result in this section, which is essentially a multivariate extension of Theorem 3.1 from Marchand [40].

Theorem 2. *Suppose that $n > 2$ and let $s_1 \geq \cdots \geq s_{d\wedge n} > 0$ denote the nonzero (with probability 1) eigenvalues of $n^{-1}X^TX$. Let $\tau \geq 0$.*

(a) *If $d \leq n$ and $\boldsymbol{\beta} \in S^{d-1}(\tau)$, then*

$$R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} \leq R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} \leq R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} + \frac{1}{d}E\left[\frac{s_1}{s_d}\operatorname{tr}\left\{\left(X^TX + \frac{d}{\tau^2}I_d\right)^{-1}\right\}\right].$$

(b) *If $d > n$ and $\boldsymbol{\beta} \in S^{d-1}(\tau)$, then*

$$R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} \leq R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} \leq R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} + \frac{1}{n}E\left[\frac{s_1}{s_n}\operatorname{tr}\left\{\left(XX^T + \frac{d}{\tau^2}I_n\right)^{-1}\right\}\right]$$
$$+ \frac{2(d-n)}{\tau^2(n-2)}E\left[\operatorname{tr}\left\{\left(XX^T + \frac{d}{\tau^2}I_n\right)^{-2}\right\}\right].$$

Theorem 2 is proved in Appendix B. The bound $R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} \leq R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}$ follows immediately from Proposition 3(b) and Corollary 1. Proving the required upper bounds on $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}$ (which, by Proposition 3(b), are equivalent to lower bounds on $r(\tau)$) is fairly complex and involves transforming the linear model into an equivalent sequence model, along with the application of classical information identities (Brown [14]) and inequalities (Stam [47]). In the remainder of this section, we discuss some of the implications of Theorem 2.

Asymptotically, Theorem 2 is primarily significant for settings where $d/n \to \rho \in (0, \infty)$. If $d/n \to \rho \in (0, \infty) \setminus \{1\}$, then the upper bounds in Theorem 2 are $\mathrm{O}(n^{-1})$ and $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} \approx R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} = r(\tau)$, where $\boldsymbol{\beta} \in S^{d-1}(\tau)$; by Corollary 2, we can further conclude that $r(\tau) \approx R_r(\tau, d/n)$. The case where $d/n \to 1$ is somewhat problematic, because then $E(s_d^{-1}) \to \infty$; however, some conclusions can be made in this case by continuity arguments, for example, Corollary 3(b) below.

Proposition 4. *Suppose that $0 < \rho_- \leq d/n \leq \rho_+ < \infty$ for some fixed constants $\rho_-, \rho_+ \in \mathbb{R}$ and that $0 < \rho_- < \rho_+ < 1$ or $1 < \rho_- < \rho_+ < \infty$. If $|n - d| > 5$, then*

$$\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} \left|R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\}\right| = \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} \left|R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - r(\tau)\right|$$
$$= \mathrm{O}\left(\frac{\tau^2}{\tau^2 + 1}n^{-1}\right).$$

Corollary 3. *Let $R_r(\tau, \rho)$ be the asymptotic risk of the ridge estimator defined in Corollary 2.*

(a) *If $\rho \in (0, \infty) \setminus \{1\}$, then*

$$\lim_{d/n \to \rho} \sup_{0 \leq \tau} \big| R_r(\tau, d/n) - r(\tau) \big| = \lim_{d/n \to \rho} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} \big| R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big), \boldsymbol{\beta}\big\} - r\big(\|\boldsymbol{\beta}\|\big) \big| = 0.$$

(b) *If $0 \leq T < \infty$ is a fixed real number, then*

$$\lim_{d/n \to 1} \sup_{0 \leq \tau \leq T} \big| R_r(\tau, d/n) - r(\tau) \big| = \lim_{d/n \to 1} \sup_{\substack{\boldsymbol{\beta} \in \mathbb{R}^d; \\ 0 \leq \|\boldsymbol{\beta}\| \leq T}} \big| R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big), \boldsymbol{\beta}\big\} - r\big(\|\boldsymbol{\beta}\|\big) \big| = 0.$$

Proposition 4 follows directly from Theorem 2 and Lemma C.2 (found in Appendix C). Corollary 3(a) follows immediately from Proposition 4 and Corollary 2(a). Corollary 3(b) may be proved similarly to part (a), while making use of the inequality $r_{d,n-k}(\tau) \leq r_{d,n}(\tau)$ for integers $0 \leq k < n$ in order to avoid issues around $d/n \approx 1$. Corollary 3 implies that if $d/n \to \rho \in (0, \infty)$, then the minimax risk $r(\tau)$ is asymptotically equivalent to the asymptotic risk of the oracle ridge estimator and that the oracle ridge estimator is asymptotically minimax.

Corollary 3 also provides the means for relating the minimax problem over ℓ^2-spheres (4) to the minimax problem over ℓ^2-balls (5). Since $S^{d-1}(\tau) \subseteq B_d(\tau)$, we have $r(\tau) \leq \bar{r}(\tau)$. Furthermore, one easily checks that

$$\sup_{\boldsymbol{\beta} \in B_d(\tau)} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} = \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\}.$$

Thus, if $d/n \to \rho \in (0, \infty)$, then

$$r(\tau) \leq \bar{r}(\tau) \leq \sup_{\boldsymbol{\beta} \in B_d(\tau)} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} = \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} \to r(\tau). \tag{12}$$

It follows that if $d/n \to \rho \in (0, \infty)$, then the minimax risk over $S^{d-1}(\tau)$ is equivalent to the minimax risk over $B_d(\tau)$ and that the ridge estimator $\hat{\boldsymbol{\beta}}_r(\tau)$ is asymptotically minimax for both problems.

When $d/n \to 0$ or ∞, asymptotics for the minimax risk $r(\tau)$ are more straightforward. The following proposition summarizes the behavior of $r(\tau)$ in these settings.

Proposition 5.

(a) $[d/n \to 0]$ *Let $R_r^0(\tau, \rho)$ be the risk function (9). Then*

$$\lim_{\substack{d/n \to 0 \\ d \to \infty}} \sup_{\tau \geq 0} \left| \frac{R_r^0(\tau, d/n)}{r(\tau)} - 1 \right| = 0.$$

(b) $[d/n \to \infty]$ *Let $0 < T < \infty$ be fixed. Then*

$$\lim_{d/n \to \infty} \sup_{0 \leq \tau \leq T} \big| r(\tau) - \tau^2 \big| = 0.$$

Proposition 5(a) is a straightforward consequence of Theorem 2, Proposition 2, and Lemma C.2. Proposition 5(b) follows from general properties of orthogonally equivariant estimators; in particular, one can check that if $d \geq n$, then

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq \frac{d-n}{d}\|\boldsymbol{\beta}\|^2$$

for all orthogonally equivariant estimators $\hat{\boldsymbol{\beta}}$.

Proposition 5 gives precise asymptotics for $r(\tau)$ when $d \to \infty$ and $d/n \to 0$ or ∞. While Proposition 5 does not directly reference the ridge estimator, combined with Proposition 2 it implies that $\hat{\boldsymbol{\beta}}_r(\tau)$ is asymptotically optimal for the minimax problem (4) when $d \to \infty$ and $d/n \to 0$ or ∞. Note that the null estimator $\hat{\boldsymbol{\beta}}_{\text{null}} = 0$ is also asymptotically optimal for (4) when $d/n \to \infty$. We point out that the condition $d \to \infty$ in Proposition 5(a) appears to be necessary, as it drives the approximation $\pi_{S^{d-1}(\tau)} \approx N(0, \tau^2/d\, I_d)$ underlying Theorem 2.

4. An adaptive ridge estimator

To this point, we have focused on the oracle ridge estimator $\hat{\boldsymbol{\beta}}_r(\tau)$, where $\tau = \|\boldsymbol{\beta}\|$ is the signal strength. Typically, τ is unknown and, consequently, $\hat{\boldsymbol{\beta}}_r(\tau)$ is non-implementable. A natural strategy is to replace τ with an estimate, $\hat{\tau}$.

Define

$$\hat{\tau}^2 = \max\left\{\frac{1}{n}\|\mathbf{y}\|^2 - 1, 0\right\} \tag{13}$$

and define the adaptive ridge estimator

$$\check{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}_r(\hat{\tau}).$$

Observe that $\check{\boldsymbol{\beta}}_r \in \mathscr{E}$ is orthogonally equivariant. One can check that $\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} E_{\boldsymbol{\beta}}(\hat{\tau}/\tau - 1)^2 \to 0$ whenever $n \to \infty$ (see Lemma C.5); thus, $\hat{\tau}$ is a reasonable estimator for τ. The next result relates the risk of the adaptive ridge estimator $\check{\boldsymbol{\beta}}_r$ to that of the oracle ridge estimator. It is proved in Appendix A.

Theorem 3. *Suppose that $0 < \rho_- < d/n < \rho_+ < \infty$, where $\rho_-, \rho_+ \in \mathbb{R}$ are fixed constants satisfying $0 < \rho_- < \rho_+ < 1$ or $1 < \rho_- < \rho_+ < \infty$. Also suppose that $|n - d| > 9$ and $n > 8$. Then*

$$\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} \left|R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta}) - R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}\right| = \mathrm{O}\left(\frac{1}{\tau^2 + 1} n^{-1/2}\right) \tag{14}$$

and

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^d} \left|R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta}) - R_r\big(\|\boldsymbol{\beta}\|, d/n\big)\right| = \mathrm{O}\big(n^{-1/2}\big), \tag{15}$$

where $R_r(\tau, \rho)$ is the asymptotic risk of the oracle ridge estimator defined in Corollary 2.

Theorem 3 implies that if $d/n \to \rho \in (0, \infty) \setminus \{1\}$, then the risk of the adaptive ridge estimator converges uniformly to that of the oracle ridge estimator and its asymptotic risk is given explicitly by $R_r(\tau, \rho)$. If $d/n \to \rho \in (0, \infty) \setminus \{1\}$ and $\tau^2 = \|\boldsymbol{\beta}\|^2 \gg n^{-1/2}$, then it follows from Theorem 3 that $R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta})/R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} \to 1$. On the other hand, if $d/n \to \rho \in (0, \infty) \setminus \{1\}$ and $\tau^2 = \mathrm{O}(n^{-1/2})$, then $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} = \mathrm{O}(n^{-1/2})$ and the limit of $R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta})/R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}$ does not follow readily from Theorem 3. In other words, the effectiveness of the adaptive ridge estimator is less clear when $\tau^2 = \|\boldsymbol{\beta}\|^2$ is very small.

If $d/n \to 0$ or $d/n \to 1$, then results similar to Theorem 3 may be obtained for the adaptive ridge estimator, but the results are more delicate; results for $d/n \to \infty$ are, in a sense, unnecessary because the oracle ridge estimator is equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{null}}$ in this setting. If $d/n \to 0$, then the relevant quantity is the risk ratio $R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta})/R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\}$, rather than the risk difference considered in Theorem 3, and one must carefully track the magnitude of $\tau^2 = \|\boldsymbol{\beta}\|^2$ relative to d/n. Ultimately, however, when $d/n \to 0$ the message is the same as the case where $d/n \to \rho \in (0, \infty) \setminus \{1\}$: If τ^2 is not too small, then the adaptive ridge estimator performs nearly as well as the oracle ridge estimator. If $d/n \to 1$, then the rate in (14) may be different, depending on τ^2 and the magnitude of $|d/n - 1|$, for example, Bai *et al.* [2].

4.1. Adaptive minimax estimation

Theorem 3 compares the risk of the adaptive ridge estimator to that of the oracle ridge estimator. The next result, which follows immediately from Theorem 3 and Proposition 4, compares the risk of the adaptive ridge estimator to $r(\tau)$.

Proposition 6. *Suppose that $\rho_-, \rho_+ \in \mathbb{R}$ are fixed constants satisfying $0 < \rho_- < \rho_+ < 1$ or $1 < \rho_- < \rho_+ < \infty$. Suppose further that $0 < \rho_- \leq d/n \leq \rho_+ < \infty$. If $|n - d| > 9$ and $n > 8$, then*

$$\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} \left| R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta}) - r(\tau) \right| = \mathrm{O}\left(\frac{\tau^2}{\tau^2 + 1} n^{-1} \right) + \mathrm{O}\left(\frac{1}{\tau^2 + 1} n^{-1/2} \right). \tag{16}$$

Combined with Proposition 4, Proposition 6 implies that if $d/n \to \rho \in (0, \infty) \setminus \{1\}$, then $\check{\boldsymbol{\beta}}_r$ is adaptive asymptotic minimax over spheres $S^{d-1}(\tau)$, provided $\tau^2 \gg n^{-1/2}$.

4.2. Equivariance

In Section 3.1, we discussed connections between the minimax problem (4) and equivariance. Previously in this section, we noted that the adaptive ridge estimator $\check{\boldsymbol{\beta}}_r$ is orthogonally equivariant and adaptive asymptotic minimax over spheres $S^{d-1}(\tau)$. The following is an asymptotic optimality result for $\check{\boldsymbol{\beta}}_r$, which pertains to the class of orthogonally equivariant estimators $\mathscr{E}$.

Proposition 7. *Suppose that $\rho \in (0, \infty) \setminus \{1\}$. Then*

$$\lim_{d/n \to \rho} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} \left| R(\check{\boldsymbol{\beta}}_r, \boldsymbol{\beta}) - \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \right| = 0.$$

By Proposition 3, $\inf_{\hat{\boldsymbol{\beta}}\in\mathscr{E}} R(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) = r(\|\boldsymbol{\beta}\|)$. Thus, Proposition 7 is a direct consequence of Proposition 6. Proposition 7 implies that if $d/n \to \rho \in (0,\infty)\setminus\{1\}$, then the adaptive ridge estimator $\breve{\boldsymbol{\beta}}_r$ is asymptotically optimal among all orthogonally equivariant estimators. Note that the caveats discussed after the statement of Theorem 3 relating to small $\|\boldsymbol{\beta}\|$ also apply to Proposition 7. More specifically, if $\|\boldsymbol{\beta}\| = \mathrm{O}(n^{-1/2})$, then the ratio $R(\breve{\boldsymbol{\beta}}_r,\boldsymbol{\beta})/\{\inf_{\hat{\boldsymbol{\beta}}\in\mathscr{E}} R(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})\}$ is more relevant than the risk difference considered in Proposition 7 and the precise asymptotic behavior of this ratio is less clear.

Appendix A

This appendix contains proofs of results stated in the main text, with the exception of Theorem 2; a proof of Theorem 2 may be found in Appendix B.

Proof of Proposition 1. Fix $t \in [0,\infty]$ and suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$. Then

$$\begin{aligned} R\{\hat{\boldsymbol{\beta}}_r(t),\boldsymbol{\beta}\} &= E_{\boldsymbol{\beta}}\{\|\hat{\boldsymbol{\beta}}_r(t)-\boldsymbol{\beta}\|^2\} \\ &= E\{\|d(t^2X^TX+dI_d)^{-1}\boldsymbol{\beta} - t^2(t^2X^TX+dI_d)^{-1}X^T\boldsymbol{\varepsilon}\|^2\} \\ &= E\{\|d(t^2X^TX+dI_d)^{-1}\boldsymbol{\beta}\|^2\} + E\{\|t^2(t^2X^TX+dI_d)^{-1}X^T\boldsymbol{\varepsilon}\|^2\}. \end{aligned} \tag{17}$$

Since X is orthogonally invariant (i.e., X and XU have the same distribution for any $U \in \mathcal{O}(d)$), it follows that

$$\begin{aligned} E\{\|d(t^2X^TX+dI_d)^{-1}\boldsymbol{\beta}\|^2\} &= d^2E\{\boldsymbol{\beta}^T(t^2X^TX+dI_d)^{-2}\boldsymbol{\beta}\} \\ &= d^2\tau^2E\{\mathbf{e}_k^T(t^2X^TX+dI_d)^{-2}\mathbf{e}_k\}, \end{aligned}$$

where $\mathbf{e}_k = (0,\dots,0,1,0,\dots,0)^T \in \mathbb{R}^d$ is the kth standard basis vector. Summing over $k = 1,\dots,d$ above and dividing by d, we obtain

$$E\{\|d(t^2X^TX+dI_d)^{-1}\boldsymbol{\beta}\|^2\} = d\tau^2E[\mathrm{tr}\{(t^2X^TX+dI_d)^{-2}\}]. \tag{18}$$

Additionally, it is clear that

$$E\{\|t^2(t^2X^TX+dI_d)^{-1}X^T\boldsymbol{\varepsilon}\|^2\} = E[\mathrm{tr}\{t^4(t^2X^TX+dI_d)^{-2}X^TX\}].$$

Combining this with (17) and (18) yields

$$\begin{aligned} R\{\hat{\boldsymbol{\beta}}_r(t),\boldsymbol{\beta}\} &= d\tau^2E[\mathrm{tr}\{(t^2X^TX+dI_d)^{-2}\}] + E[\mathrm{tr}\{t^4(t^2X^TX+dI_d)^{-2}X^TX\}] \\ &= E[\mathrm{tr}\{(t^2X^TX+dI_d)^{-2}(t^4X^TX+d\tau^2I_d)\}]. \end{aligned}$$

Now let $s_1 \geq \cdots \geq s_d \geq 0$ denote the eigenvalues of $n^{-1}X^TX$. Then

$$R\{\hat{\boldsymbol{\beta}}_r(t), \boldsymbol{\beta}\} = E\left\{\sum_{j=1}^{d} \frac{t^4 n s_j + d\tau^2}{(t^2 n s_j + d)^2}\right\}$$
$$= E\left[\sum_{j=1}^{d}\left\{\frac{\tau^2}{\tau^2 n s_j + d} + \frac{d n s_j (\tau^2 - t^2)^2}{(t^2 n s_j + d)^2(\tau^2 n s_j + d)}\right\}\right].$$

Clearly, the right-hand side above is minimized by taking $t = \tau$ and $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} = E[\mathrm{tr}\{(X^TX + d/\tau^2 I_d)^{-1}\}]$. □

Proof of Theorem 1. Suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$ and let $\mathbb{F}_{n,d}$ be the empirical cumulative distribution function of the eigenvalues of $n^{-1}X^TX$. Using integration by parts, for $c \geq 0$,

$$\frac{n}{d}\mathrm{tr}\{(X^TX + d/\tau^2 I_d)^{-1}\} = \int_0^\infty \frac{1}{s + d/(n\tau^2)}\,\mathrm{d}\mathbb{F}_{n,d}(s)$$
$$= \int_0^c \frac{1}{s + d/(n\tau^2)}\,\mathrm{d}\mathbb{F}_{n,d}(s) + \frac{1}{c + d/(n\tau^2)}\{1 - \mathbb{F}_{n,d}(c)\} \quad (19)$$
$$- \int_c^\infty \frac{1}{\{s + d/(n\tau^2)\}^2}\{1 - \mathbb{F}_{n,d}(s)\}\,\mathrm{d}s.$$

Similarly,

$$m_{d/n}\{-d/(n\tau^2)\} = \int_0^c \frac{1}{s + d/(n\tau^2)}\,\mathrm{d}F_{d/n}(s) + \frac{1}{c + d/(n\tau^2)}\{1 - F_{d/n}(c)\}$$
$$- \int_c^\infty \frac{1}{\{s + d/(n\tau^2)\}^2}\{1 - F_{d/n}(s)\}\,\mathrm{d}s. \quad (20)$$

Now let $\Delta = |R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - (d/n)m_{d/n}\{-d/(n\tau^2)\}|$. Taking $c = 0$ in (19) and (20) implies

$$\Delta \leq \frac{d}{n}\int_0^\infty \frac{1}{\{s + d/(n\tau^2)\}^2}\left|E\{\mathbb{F}_{n,d}(s)\} - F_{d/n}(s)\right|\mathrm{d}s$$
$$\leq \|\boldsymbol{\beta}\|^2 \sup_{s\geq 0}\left|E\{\mathbb{F}_{n,d}(s)\} - F_{d/n}(s)\right|,$$

where we have used the fact that $\mathbb{F}_{n,d}(0) = F_{d/n}(0) = \max\{1 - n/d, 0\}$, with probability 1. Thus, it follows from Theorem 1.1 of Bai *et al.* [2] (see equation (8) in Section 2.2 above) that

$$\Delta = \begin{cases} O(\tau^2 n^{-1/2}) & \text{if } 0 < \rho_- < \rho_+ < 1 \text{ or } 1 < \rho_- < \rho_+ < \infty, \\ O(\tau^2 n^{-1/8}) & \text{if } 0 < \rho_- < 1 < \rho_+ < \infty. \end{cases}$$

Part (b) of Theorem 1 follows immediately.

To prove Theorem 1(a) we show that, in fact, $\Delta = \mathrm{O}(n^{-1/2})$ if $0 < \rho_- < \rho_+ < 1$ or $1 < \rho_- < \rho_+ < \infty$. First, suppose that $0 < \rho_- < \rho_+ < 1$. Then, for $0 < c < (1 - \sqrt{d/n})^2$,

$$m_{d/n}\biggl(-\frac{d}{n\tau^2}\biggr) = \frac{1}{c + d/(n\tau^2)} - \int_c^\infty \frac{1}{\{s + d/(n\tau^2)\}^2}\bigl\{1 - F_{d/n}(s)\bigr\}\,\mathrm{d}s$$

and

$$\begin{aligned}
\frac{n}{d}\Delta &\le E\biggl\{\int_0^c \frac{1}{s + d/(n\tau^2)}\,\mathrm{d}\mathbb{F}_{n,d}(s)\biggr\} + \frac{1}{c + d/(n\tau^2)}E\bigl\{\mathbb{F}_{n,d}(c)\bigr\} \\
&\quad + \biggl|\int_c^\infty \frac{1}{\{s + d/(n\tau^2)\}^2}\bigl[E\bigl\{\mathbb{F}_{n,d}(s)\bigr\} - F_{d/n}(s)\bigr]\,\mathrm{d}s\biggr| \\
&\le E\biggl\{\int_0^c s^{-1}\,\mathrm{d}\mathbb{F}_{n,d}(s)\biggr\} + \frac{1}{c + d/(n\tau^2)}E\bigl\{\mathbb{F}_{n,d}(c)\bigr\} \\
&\quad + \frac{1}{c + d/(n\tau^2)}\sup_{s\ge c}\bigl|E\bigl\{\mathbb{F}_{n,d}(s)\bigr\} - F_{d/n}(s)\bigr| \\
&\le E\bigl[s_d^{-1}\mathbf{1}_{\{s_d < c\}}\bigr] + \frac{1}{c + d/(n\tau^2)}P(s_d < c) \\
&\quad + \frac{1}{c + d/(n\tau^2)}\sup_{s\ge c}\bigl|E\bigl\{\mathbb{F}_{n,d}(s)\bigr\} - F_{d/n}(s)\bigr| \\
&\le \bigl\{E\bigl(s_d^{-2}\bigr)\bigr\}^{1/2}P(s_d < c)^{1/2} + c^{-1}P(s_d < c) + c^{-1}\sup_{s\ge c}\bigl|E\bigl\{\mathbb{F}_{n,d}(s)\bigr\} - F_{d/n}(s)\bigr|,
\end{aligned}$$

where $s_d \ge 0$ is the smallest eigenvalue of $n^{-1}X^TX$, $\mathbf{1}_D$ is the indicator function of the event D, and $P(\cdot)$ denotes the probability measure induced by the joint distribution of $(X, \boldsymbol{\varepsilon})$. We bound the first two terms and the last term on right-hand side above separately. Bounding the first two terms relies on a result of Davidson and Szarek [22]. Their Theorem II.13, which is a consequence of concentration of measure, implies that

$$P(s_d \le c) \le \exp\biggl\{-\frac{n(1 - \sqrt{d/n})^2}{2}\biggl(1 - \frac{c^{1/2}}{1 - \sqrt{d/n}}\biggr)^2\biggr\}, \tag{21}$$

provided $c \le 1 - \sqrt{d/n}$. Additionally, Lemma C.2 in Appendix C implies that $E(s_d^{-2}) = \mathrm{O}(1)$ if $n - d > 5$. Taking $c = (1 - \sqrt{d/n})^2/2$, it follows that

$$\bigl\{E\bigl(s_d^{-2}\bigr)\bigr\}^{1/2}P(s_d < c)^{1/2} + c^{-1}P(s_d < c) = \mathrm{O}\bigl(n^{-1/2}\bigr)$$

(in fact, we can conclude that the quantities on the left above decay exponentially, but this is not required for the current result). It now follows from Theorem 1.1 of Bai *et al.* [2] that $\Delta = \mathrm{O}(n^{-1/2})$. For the case where $1 < \rho_- < \rho_+ < \infty$, we note that the same argument as above may be applied, except that both $\mathbb{F}_{n,d}(s)$ and $F_{d/n}(s)$ have a mass of weight $(d - n)/d$ at 0, which cancel. Theorem 1(a) follows. □

Proof of Proposition 2. Proposition 2(b) follows directly from Proposition 1. Part (a) follows from two applications of Jensen's inequality. If $d+1<n$, then

$$\begin{aligned} R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big),\boldsymbol{\beta}\big\} &= E\big[\mathrm{tr}\big\{\big(X^TX+d/\|\boldsymbol{\beta}\|^2I_d\big)^{-1}\big\}\big] \\ &\geq d\bigg[\frac{1}{d}E\big\{\mathrm{tr}\big(X^TX\big)\big\}+\frac{d}{\|\boldsymbol{\beta}\|^2}\bigg]^{-1} \\ &= \frac{\|\boldsymbol{\beta}\|^2d/n}{\|\boldsymbol{\beta}\|^2+d/n} \\ &= R_r^0\big(\|\boldsymbol{\beta}\|,d/n\big) \end{aligned}$$

and, since $E[\mathrm{tr}\{(X^TX)^{-1}\}]=d/(n-d-1)$ (Problem 3.6 of Muirhead [41]),

$$\begin{aligned} R\big\{\hat{\boldsymbol{\beta}}_r\big(\|\boldsymbol{\beta}\|\big),\boldsymbol{\beta}\big\} &= E\big[\mathrm{tr}\big\{\big(X^TX+d/\|\boldsymbol{\beta}\|^2I_d\big)^{-1}\big\}\big] \\ &\leq \frac{E[\mathrm{tr}\{(X^TX)^{-1}\}]}{1+(1/\|\boldsymbol{\beta}\|^2)E[\mathrm{tr}\{(X^TX)^{-1}\}]} \\ &= \frac{\|\boldsymbol{\beta}\|^2d/(n-d-1)}{\|\boldsymbol{\beta}\|^2+d/(n-d-1)} \\ &= R_r^0\big\{\|\boldsymbol{\beta}\|,d/(n-d-1)\big\}. \end{aligned}$$

Thus, $R_r^0\{\|\boldsymbol{\beta}\|,d/(n-d-1)\}\leq R\{\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|),\boldsymbol{\beta}\}\leq R_r^0(\|\boldsymbol{\beta}\|,d/n)$. It follows that if $d/n\to 0$, then

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^d}\bigg|\frac{R\{\hat{\boldsymbol{\beta}}_r(\|\boldsymbol{\beta}\|),\boldsymbol{\beta}\}}{R_r^0(\|\boldsymbol{\beta}\|,d/n)}-1\bigg|\to 0. \qquad \square$$

Proof of Proposition 3. Suppose that $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(\mathbf{y},X)\in\mathscr{E}$ and that $\boldsymbol{\beta}\in S^{d-1}(\tau)$. Let $\mathbf{e}_1=(1,0,\ldots,0)\in\mathbb{R}^d$ denote the first standard basis vector and let $U\in\mathcal{O}(d)$ satisfy $\boldsymbol{\beta}=\tau U\mathbf{e}_1$. Then, since $\hat{\boldsymbol{\beta}}\in\mathscr{E}$ and $(X,\boldsymbol{\varepsilon})$ has the same distribution as $(XU,\boldsymbol{\varepsilon})$,

$$\begin{aligned} R(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}}\big(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2\big) \\ &= E_{\boldsymbol{\beta}}\big(\big\|U^T\hat{\boldsymbol{\beta}}(\mathbf{y},X)-\tau\mathbf{e}_1\big\|^2\big) \\ &= E_{\boldsymbol{\beta}}\big(\big\|\hat{\boldsymbol{\beta}}(\mathbf{y},XU)-\tau\mathbf{e}_1\big\|^2\big) \\ &= E\big(\big\|\hat{\boldsymbol{\beta}}(XU\tau\mathbf{e}_1+\boldsymbol{\varepsilon},XU)-\tau\mathbf{e}_1\big\|^2\big) \\ &= E\big(\big\|\hat{\boldsymbol{\beta}}(X\tau\mathbf{e}_1+\boldsymbol{\varepsilon},X)-\tau\mathbf{e}_1\big\|^2\big) \\ &= E_{\tau\mathbf{e}_1}\big(\|\hat{\boldsymbol{\beta}}-\tau\mathbf{e}_1\|^2\big) \\ &= R(\hat{\boldsymbol{\beta}},\tau\mathbf{e}_1). \end{aligned}$$

Part (a) of the proposition follows.

To prove part (b), we first show that

$$r(\tau) = \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}). \tag{22}$$

Given an estimator $\hat{\boldsymbol{\beta}}$ (not necessarily orthogonally equivariant), define

$$\hat{\boldsymbol{\beta}}_{\mathcal{O}}(\mathbf{y}, X) = \int_{\mathcal{O}(d)} U \hat{\boldsymbol{\beta}}(\mathbf{y}, XU) \, \mathrm{d}\pi_{\mathcal{O}(d)}(U),$$

where $\pi_{\mathcal{O}(d)}$ is the uniform (Haar) measure on $\mathcal{O}(d)$. Then $\hat{\boldsymbol{\beta}} \in \mathscr{E}$ and, since X and XU have the same distribution for any $U \in \mathcal{O}(d)$,

$$\begin{aligned}
\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}_{\mathcal{O}}, \boldsymbol{\beta}) &= \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} E_{\boldsymbol{\beta}} \left\{ \left\| \int_{\mathcal{O}(d)} U \hat{\boldsymbol{\beta}}(\mathbf{y}, XU) \, \mathrm{d}\pi_{\mathcal{O}(d)}(U) - \boldsymbol{\beta} \right\|^2 \right\} \\
&\leq \int_{\mathcal{O}(d)} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} E \left\{ \left\| U \hat{\boldsymbol{\beta}}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, XU) - \boldsymbol{\beta} \right\|^2 \right\} \mathrm{d}\pi_{\mathcal{O}(d)}(U) \\
&= \int_{\mathcal{O}(d)} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} E \left\{ \left\| \hat{\boldsymbol{\beta}}\left(XU^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}, X\right) - U^T \boldsymbol{\beta} \right\|^2 \right\} \mathrm{d}\pi_{\mathcal{O}(d)}(U) \\
&\leq \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}).
\end{aligned}$$

The identity (22) follows. Thus, by part (a) and the fact that $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)} \in \mathscr{E}$,

$$\begin{aligned}
r(\tau) &= \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} \sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \\
&= \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1) \\
&= \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} E_{\pi_{S^{d-1}(\tau)}}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \\
&= E_{\pi_{S^{d-1}(\tau)}}\left\{\|\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)} - \boldsymbol{\beta}\|^2\right\} \\
&= r_B(\tau),
\end{aligned}$$

which completes the proof of the proposition. □

Proof of Theorem 3. Suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$. It is clear that (15) follows from (14) and Theorem 1. To prove (14), consider the risk decomposition of the oracle and adaptive ridge

estimators

$$
\begin{aligned}
R\{\hat{\boldsymbol{\beta}}_r(\tau),\boldsymbol{\beta}\} &= \left(\frac{d}{n}\right)^2 E\left\{\left\|\left(\frac{\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}\boldsymbol{\beta}\right\|^2\right\} \\
&\quad + \frac{1}{n^2}E\left\{\left\|\tau^2\left(\frac{\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}X^T\boldsymbol{\varepsilon}\right\|^2\right\}, \\
R(\check{\boldsymbol{\beta}}_r,\boldsymbol{\beta}) &= \left(\frac{d}{n}\right)^2 E_{\boldsymbol{\beta}}\left\{\left\|\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}\boldsymbol{\beta}\right\|^2\right\} \\
&\quad - 2\frac{d}{n^2}E_{\boldsymbol{\beta}}\left\{\hat{\tau}^2\boldsymbol{\varepsilon}^TX\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-2}\boldsymbol{\beta}\right\} \\
&\quad + \frac{1}{n^2}E_{\boldsymbol{\beta}}\left\{\left\|\hat{\tau}^2\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}X^T\boldsymbol{\varepsilon}\right\|^2\right\}.
\end{aligned}
$$

The triangle inequality implies

$$
\left|R\{\hat{\boldsymbol{\beta}}_r(\tau),\boldsymbol{\beta}\}-R(\check{\boldsymbol{\beta}}_r,\tau)\right| \le \left|E_{\boldsymbol{\beta}}(H_1)\right| + \left|E_{\boldsymbol{\beta}}(H_2)\right| + 2\left|E_{\boldsymbol{\beta}}(H_3)\right|, \tag{23}
$$

where

$$
\begin{aligned}
H_1 &= \left(\frac{d}{n}\right)^2\left\{\left\|\left(\frac{\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}\boldsymbol{\beta}\right\|^2 - \left\|\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}\boldsymbol{\beta}\right\|^2\right\}, \\
H_2 &= \frac{1}{n^2}\left\{\left\|\tau^2\left(\frac{\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}X^T\boldsymbol{\varepsilon}\right\|^2 - \left\|\hat{\tau}^2\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-1}X^T\boldsymbol{\varepsilon}\right\|^2\right\}, \\
H_3 &= \frac{d}{n^2}\hat{\tau}^2\boldsymbol{\varepsilon}^TX\left(\frac{\hat{\tau}^2}{n}X^TX+\frac{d}{n}I_d\right)^{-2}\boldsymbol{\beta}.
\end{aligned}
$$

To prove the theorem, we bound the terms $|E_{\boldsymbol{\beta}}(H_1)|$, $|E_{\boldsymbol{\beta}}(H_2)|$, and $|E_{\boldsymbol{\beta}}(H_3)|$ separately.

Let $s_1 \ge \cdots \ge s_d \ge 0$ denote the ordered eigenvalues of $n^{-1}X^TX$ and let $U \in \mathcal{O}(d)$ be a $d \times d$ orthogonal matrix such that $S = n^{-1}U^TX^TXU$ is diagonal. Additionally, let $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1,\ldots,\tilde{\beta}_d)^T = U^T\boldsymbol{\beta}$ and let $\tilde{\boldsymbol{\delta}} = (\tilde{\delta}_1,\ldots,\tilde{\delta}_d)^T = U^T(X^TX)^{-1/2}X^T\boldsymbol{\varepsilon}$, where $(X^TX)^{-1/2}$ denotes the Moore–Penrose pseudoinverse of $(X^TX)^{1/2}$ if X^TX is not invertible. Then

$$
\begin{aligned}
|H_1| &= \left(\frac{d}{n}\right)^2\left|\sum_{j=1}^d\left\{\frac{\tilde{\beta}_j^2}{(\hat{\tau}^2s_j+d/n)^2}-\frac{\tilde{\beta}_j^2}{(\tau^2s_j+d/n)^2}\right\}\right| \\
&= \left(\frac{d}{n}\right)^2\left|\sum_{j=1}^d\frac{\tilde{\beta}_j^2s_j(\tau^2-\hat{\tau}^2)}{(\hat{\tau}^2s_j+d/n)(\tau^2s_j+d/n)}\left(\frac{1}{\hat{\tau}^2s_j+d/n}+\frac{1}{\tau^2s_j+d/n}\right)\right|.
\end{aligned}
$$

Since $(ax+b)^{-1} \le (a+b)^{-1}\max\{x^{-1},1\}$ for $a,b,x \ge 0$,

$$
\begin{aligned}
|H_1| &\le \left(\frac{d}{n}\right)^2 \sum_{j=1}^{d\wedge n}\left\{\frac{\tilde\beta_j^2|\tau^2-\hat\tau^2|}{(\hat\tau^2+d/n)(\tau^2+d/n)}\left(\frac{1}{\hat\tau^2+d/n}+\frac{1}{\tau^2+d/n}\right)\left(\frac{1}{s_j^2}+s_j\right)\right\} \\
&\le \left(\frac{d}{n}\right)^2 \frac{|\tau^2-\hat\tau^2|}{\hat\tau^2+d/n}\left(\frac{1}{\hat\tau^2+d/n}+\frac{1}{\tau^2+d/n}\right)\left(\frac{1}{s_{d\wedge n}^2}+s_1\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
|H_2| &= \frac{1}{n}\left|\sum_{j=1}^{d}\left\{\frac{\hat\tau^4 s_j\tilde\delta_j^2}{(\hat\tau^2 s_j+d/n)^2}-\frac{\tau^4 s_j\tilde\delta_j^2}{(\tau^2 s_j+d/n)^2}\right\}\right| \\
&= \frac{1}{n}\left|\sum_{j=1}^{d}\frac{(d/n)\tilde\delta_j^2 s_j(\hat\tau^2-\tau^2)}{(\hat\tau^2 s_j+d/n)(\tau^2 s_j+d/n)}\left(\frac{\hat\tau^2}{\hat\tau^2 s_j+d/n}+\frac{\tau^2}{\tau^2 s_j+d/n}\right)\right| \\
&\le \frac{1}{n}\sum_{j=1}^{d\wedge n}\frac{(d/n)\tilde\delta_j^2|\hat\tau^2-\tau^2|}{(\hat\tau^2+d/n)(\tau^2+d/n)}\left(\frac{1}{s_j}+s_j\right) \\
&\le \frac{d}{n^2}\|\tilde{\boldsymbol\delta}\|^2\frac{|\hat\tau^2-\tau^2|}{(\hat\tau^2+d/n)(\tau^2+d/n)}\left(\frac{1}{s_{d\wedge n}}+s_1\right).
\end{aligned}
$$

Repeated application of Hölder's inequality and Lemmas C.2, C.3 and C.5 (found in Appendix C) imply that

$$
\big|E_{\boldsymbol\beta}(H_1)\big|+\big|E_{\boldsymbol\beta}(H_2)\big| = \mathrm{O}\left(\frac{1}{\tau^2+1}n^{-1/2}\right). \tag{24}
$$

To bound $|E_{\boldsymbol\beta}(H_3)|$, we condition on X and use integration by parts (Stein's lemma, e.g., Lemma 3.6 of Tsybakov [51]):

$$
\begin{aligned}
E_{\boldsymbol\beta}(H_3) &= \frac{d}{n^2}E_{\boldsymbol\beta}\left\{\hat\tau^2\boldsymbol\varepsilon^T X\left(\frac{\hat\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-2}\boldsymbol\beta\right\} \\
&= \frac{2d}{n^3}E_{\boldsymbol\beta}\left[\mathbf{y}^T X\left(\frac{d}{n}I_d-\frac{\hat\tau^2}{n}X^TX\right)\left(\frac{\hat\tau^2}{n}X^TX+\frac{d}{n}I_d\right)^{-3}\boldsymbol\beta\mathbf{1}_{\{\|\mathbf{y}\|^2\ge n\}}\right] \\
&= \frac{2d}{n^3}E_{\boldsymbol\beta}\left[\sum_{j=1}^{d}\frac{(ns_j\tilde\beta_j+n^{1/2}s_j^{1/2}\tilde\delta_j)(d/n-\hat\tau^2 s_j)\tilde\beta_j}{(\hat\tau^2 s_j+d/n)^3}\mathbf{1}_{\{\|\mathbf{y}\|^2\ge n\}}\right].
\end{aligned}
$$

It follows that

$$\begin{aligned}\big|E_{\boldsymbol{\beta}}(H_3)\big| &\le \frac{2d}{n^3}E_{\boldsymbol{\beta}}\left\{\sum_{j=1}^d\left|\frac{(ns_j\tilde{\beta}_j + n^{1/2}s_j^{1/2}\tilde{\delta}_j)\tilde{\beta}_j}{(\hat{\tau}^2 s_j + d/n)^2}\right|\right\} \\ &\le \frac{2d}{n^2}E_{\boldsymbol{\beta}}\left\{\sum_{j=1}^d\frac{s_j\tilde{\beta}_j^2}{(\hat{\tau}^2 s_j + d/n)^2}\right\} + \frac{2d}{n^{5/2}}E_{\boldsymbol{\beta}}\left\{\sum_{j=1}^d\left|\frac{s_j^{1/2}\tilde{\delta}_j\tilde{\beta}_j}{(\hat{\tau}^2 s_j + d/n)^2}\right|\right\} \qquad (25) \\ &= \mathrm{O}\bigg(\frac{1}{\tau^2+1}n^{-1}\bigg),\end{aligned}$$

where we have used Lemmas C.2 and C.3 to obtain the last bound. The theorem follows from (23) and (25). □

Appendix B

This appendix is devoted to a proof of Theorem 2, which is fairly involved. Our first step is to show that the minimax problem (4) may be reformulated as a minimax problem for an equivalent sequence model. Ultimately, this will substantially simplify notation and allow for a direct application of results from Marchand [40] that are important for Theorem 2.

B.1. An equivalent sequence model

Let Σ be a random orthogonally invariant $m \times m$ positive semidefinite matrix with rank m, almost surely (by orthogonally invariant, we mean that Σ and $U\Sigma U^T$ have the same distribution for any $U \in \mathcal{O}(m)$). Additionally, let $\boldsymbol{\delta} \sim N(0, I_m)$ be an m-dimensional Gaussian random vector that is independent of Σ. Suppose that the observed data are $(\mathbf{w}, \Sigma)$, where

$$\mathbf{w} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta} \in \mathbb{R}^m \qquad (26)$$

and $\boldsymbol{\theta} \in \mathbb{R}^m$ is an unknown parameter.

For an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{w}, \Sigma)$, define the risk under squared error loss

$$R_{\mathrm{seq}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2,$$

where, abusing notation, the expectation $E_{\boldsymbol{\theta}}(\cdot)$ is taken with respect to $(\boldsymbol{\delta}, \Sigma)$ and the subscript $\boldsymbol{\theta}$ indicates that $\mathbf{w} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}$ (we will sometimes drop the subscript $\boldsymbol{\theta}$ in $E_{\boldsymbol{\theta}}(\cdot)$ if the integrand does not depend on $\boldsymbol{\theta}$). To distinguish $E_{\boldsymbol{\theta}}(\cdot)$ from expectations $E_{\boldsymbol{\beta}}(\cdot)$ considered elsewhere in the paper, we emphasize that all expectations considered in this section (Appendix B) refer to the sequence model (26).

B.2. Equivalence with the linear model

Most of the key concepts initially introduced in the context of the linear model (1) have analogues in the sequence model (26). Define

$$\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)} = E_{\pi_{S^{m-1}(\tau)}}(\boldsymbol{\theta}|\mathbf{w}, \Sigma),$$

to be the posterior mean of $\boldsymbol{\theta}$ under the assumption that $\boldsymbol{\theta} \sim \pi_{S^{m-1}(\tau)}$ is uniformly distributed on $S^{m-1}(\tau)$ and define

$$\hat{\boldsymbol{\theta}}_r(\tau) = E_{N(0,\tau^2/mI_m)}(\boldsymbol{\theta}|\mathbf{w}, \Sigma) = \tau^2/m\big(\Sigma + \tau^2/mI_m\big)^{-1}\mathbf{w}$$

to be the posterior mean under the assumption that $\boldsymbol{\theta} \sim N(0, \tau^2/mI_m)$ (for both of these Bayes estimators we assume that $\boldsymbol{\theta}$ is independent of $\boldsymbol{\delta}$ and Σ). The estimators $\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}(\tau)$ and $\hat{\boldsymbol{\theta}}_r(\tau)$ are analogous to the minimax estimator $\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}$ and the optimal ridge estimator $\hat{\boldsymbol{\beta}}_r(\tau)$ in the linear model, respectively. Now define the minimax risk over $S^{m-1}(\tau)$ for the sequence model

$$r_{\text{seq}}(\tau) = \inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in S(\tau)} R_{\text{seq}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}),$$

where the infimum is over all measurable estimators for $\boldsymbol{\theta}$. We have the following analogue to Proposition 3(b).

Lemma B.1. *Suppose that* $\tau \geq 0$ *and that* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in S^{m-1}(\tau)$. *Then* $R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}_1\} = R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}_2\}$ *and*

$$r_{\text{seq}}(\tau) = \sup_{\boldsymbol{\theta}\in S^{m-1}(\tau)} R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\}.$$

The proof of Lemma B.1 is essentially the same as that of Proposition 3 and is omitted. The next result gives an equivalence between the linear model (1) and the sequence model (26) when $d \leq n$.

Lemma B.2. *Suppose that* $m = d \leq n$ *and that* $\Sigma = (X^TX)^{-1}$. *Let* $\tau \geq 0$. *If* $\boldsymbol{\theta}, \boldsymbol{\beta} \in S^{m-1}(\tau)$, *then*

$$R_{\text{seq}}\big\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\big\} = R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\}$$

and

$$r_{\text{seq}}(\tau) = R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} = R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} = r(\tau).$$

Lemma B.2 follows directly upon identifying $\mathbf{w}$ with $\hat{\boldsymbol{\beta}}_{\text{ols}} = (X^TX)^{-1}X^T\mathbf{y} = \boldsymbol{\beta} + (X^TX)^{-1}X^T\boldsymbol{\varepsilon}$. Lemma B.2 implies that it suffices to consider the sequence model (26) (in particular, $R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\}$ and $R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\}$) in order to prove Theorem 2(a). Note that Lemma B.2

does not apply when $d > n$. Indeed, if $d > n$, then the usual OLS estimator is not defined (moreover, if one uses a pseudoinverse in place of $(X^TX)^{-1}$, then $(X^TX)^{-1}X^TX\boldsymbol{\beta}$ is not necessarily in $S^{d-1}(\tau)$). The case where $d > n$ is considered separately below.

B.3. Proof of Theorem 2(a)

In this section, we prove Theorem 2(a) by bounding

$$R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\}. \tag{27}$$

By Lemma B.2, this is equivalent to bounding $R\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\} - R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}(\tau), \boldsymbol{\beta}\}$. The lower bound

$$0 \leq R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} \tag{28}$$

follows immediately from Lemma B.1. Marchand [40] obtained an upper bound on (27) in the case where $\Sigma = \nu^2 I_m$ for fixed $\nu^2 > 0$ (i.e., in the Gaussian sequence model with i.i.d. errors), which is one of the keys to the proof of Theorem 2(a).

Lemma B.3 (Theorem 3.1 from Marchand [40]). *Suppose that $\Sigma = \nu^2 I_m$ for some fixed $\nu^2 > 0$ and that $\boldsymbol{\theta} \in S^{m-1}(\tau)$. Then*

$$R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} \leq \frac{1}{m}\frac{\tau^2\nu^2 m}{\tau^2 + \nu^2 m} = \frac{1}{m}R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\}.$$

Thus, in the Gaussian sequence model with i.i.d. errors, the risk of $\hat{\boldsymbol{\theta}}_r(\tau)$ is nearly as small as that of $\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}$. Marchand's result relies on somewhat delicate calculations involving modified Bessel functions (Robert [45]). A direct approach to bounding (27) for general Σ might involve attempting to mimic these calculations. However, this seems daunting (Bickel [9]). Brown's identity, which relates the risk of a Bayes estimator to the Fisher information, allows us to sidestep these calculations and apply Marchand's result directly.

Define the Fisher information of a random vector $\boldsymbol{\xi} \in \mathbb{R}^m$, with density $f_{\boldsymbol{\xi}}$ (with respect to Lebesgue measure on $\mathbb{R}^m$) by

$$I(\boldsymbol{\xi}) = \int_{\mathbb{R}^m} \frac{\nabla f_{\boldsymbol{\xi}}(\mathbf{t})\nabla f_{\boldsymbol{\xi}}(\mathbf{t})^T}{f_{\boldsymbol{\xi}}(\mathbf{t})}\,\mathrm{d}\mathbf{t},$$

where $\nabla f_{\boldsymbol{\xi}}(\mathbf{t})$ is the gradient of $f_{\boldsymbol{\xi}}(\mathbf{t})$. Brown's identity has typically been used for univariate problems or problems in the sequence model with i.i.d. Gaussian errors (Bickel [9], Brown and Gajek [16], Brown and Low [17], DasGupta [21]). The next proposition is a straightforward generalization to the correlated multivariate Gaussian setting. Its proof is based on Stein's lemma.

Lemma B.4 (Brown's identity). *Let $I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})$ denote the Fisher information of $\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}$, conditional on Σ, under the assumption that $\boldsymbol{\theta} \sim \pi_{S^{m-1}(\tau)}$ is independent of $\boldsymbol{\delta}$ and Σ. If $\boldsymbol{\theta} \in$*

$S^{m-1}(\tau)$, *then*

$$R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} = E\{\mathrm{tr}(\Sigma)\} - E\big[\mathrm{tr}\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\}\big].$$

Proof. Suppose that $\boldsymbol{\theta} \in S^{m-1}(\tau)$ and let

$$f(\mathbf{w}) = \int_{S^{m-1}(\tau)} (2\pi)^{-m/2} \det(\Sigma^{-1/2}) \mathrm{e}^{-1/2(\mathbf{w}-\boldsymbol{\theta})^T \Sigma^{-1}(\mathbf{w}-\boldsymbol{\theta})}\, \mathrm{d}\pi_{S^{m-1}(\tau)}(\boldsymbol{\theta})$$

be the density of $\mathbf{w} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}$, conditional on Σ and under the assumption that $\boldsymbol{\theta} \sim \pi_{S^{m-1}(\tau)}$. Then

$$\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)} = E_{\pi_{S^{m-1}(\tau)}}(\boldsymbol{\theta}|\mathbf{w}, \Sigma) = \mathbf{w} + \frac{\Sigma \nabla f(\mathbf{w})}{f(\mathbf{w})}.$$

It follows that

$$\begin{aligned} R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} &= E_{\pi_{S^{m-1}(\tau)}}\{\|\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)} - \boldsymbol{\theta}\|^2\} \\ &= E_{\pi_{S^{m-1}(\tau)}}\left\{\left\|\Sigma^{1/2}\boldsymbol{\delta} + \frac{\Sigma \nabla f(\mathbf{w})}{f(\mathbf{w})}\right\|^2\right\} \\ &= E\{\mathrm{tr}(\Sigma)\} + 2E_{\pi_{S^{m-1}(\tau)}}\left\{\frac{\boldsymbol{\delta}^T \Sigma^{3/2} \nabla f(\mathbf{w})}{f(\mathbf{w})}\right\} \\ &\quad + E_{\pi_{S^{m-1}(\tau)}}\left\{\frac{\nabla f(\mathbf{w})^T \Sigma^2 \nabla f(\mathbf{w})}{f(\mathbf{w})^2}\right\} \\ &= E\{\mathrm{tr}(\Sigma)\} + 2E_{\pi_{S^{m-1}(\tau)}}\left\{\frac{\boldsymbol{\delta}^T \Sigma^{3/2} \nabla f(\mathbf{w})}{f(\mathbf{w})}\right\} \\ &\quad + E\big[\mathrm{tr}\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\}\big]. \end{aligned} \tag{29}$$

By Stein's lemma (Lemma 3.6 of Tsybakov [51]),

$$\begin{aligned} E_{\pi_{S^{m-1}(\tau)}}\left\{\frac{\boldsymbol{\delta}^T \Sigma^{3/2} \nabla f(\mathbf{w})}{f(\mathbf{w})}\right\} &= E_{\pi_{S^{m-1}(\tau)}}\big[\mathrm{tr}\{\Sigma^2 \nabla^2 \log f(\mathbf{w})\}\big] \\ &= -E\big[\mathrm{tr}\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\}\big]. \end{aligned} \tag{30}$$

Brown's identity follows by combining (29) and (30). □

Using Brown's identity, Fisher information bounds may be converted to risk bounds, and vice-versa. Its usefulness in the present context springs from two observations: (i) The decomposition

$$\mathbf{w} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta} = \{\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1\} + (\Sigma - \gamma\sigma_m I_m)^{1/2}\boldsymbol{\delta}_2, \tag{31}$$

where $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \overset{\text{i.i.d.}}{\sim} N(0, I_m)$ are independent of Σ, σ_m is the smallest eigenvalue of Σ, and $0 < \gamma < 1$ is a constant; and (ii) Stam's inequality for the Fisher information of sums of independent random variables.

Lemma B.5 (Stam's inequality; this version due to Zamir [53]). *Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$ be independent random variables that are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^m$. Then*

$$\operatorname{tr}\big[\Psi I(\mathbf{u}+\mathbf{v})\big] \leq \operatorname{tr}\big[\Psi\big\{I(\mathbf{u})^{-1} + I(\mathbf{v})^{-1}\big\}^{-1}\big]$$

for all $m \times m$ positive definite matrices Ψ.

Notice in (31) that $\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1$ may be viewed as an observation from the Gaussian sequence model with i.i.d. errors, conditional on Σ. The necessary bound on (27) is obtained by piecing together Brown's identity, the decomposition (31), and Stam's inequality, so that Marchand's inequality (Lemma B.3) may be applied to $\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1$.

Lemma B.6. *Suppose that Σ has rank m with probability 1 and that $\boldsymbol{\theta} \in S^{m-1}(\tau)$. Let $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$ denote the eigenvalues of Σ. Then*

$$R_{\text{seq}}\big\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\big\} - R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} \leq E\bigg[\bigg(\frac{\sigma_1}{m\sigma_m} \wedge 1\bigg)\operatorname{tr}\big\{\big(\Sigma^{-1} + m/\tau^2 I_m\big)^{-1}\big\}\bigg].$$

Proof. Since Σ is orthogonally invariant and independent of $\boldsymbol{\delta}$,

$$\begin{aligned}
R_{\text{seq}}\big\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\big\} &= E_{\boldsymbol{\theta}}\big\{\big\|\tau^2/m\big(\Sigma + \tau^2/m I_m\big)^{-1}\mathbf{w} - \boldsymbol{\theta}\big\|^2\big\} \\
&= E\big\{\big\|\Sigma\big(\Sigma + \tau^2/m I_m\big)^{-1}\boldsymbol{\theta}\big\|^2\big\} \\
&\quad + E\big\{\big\|\tau^2/m\big(\Sigma + \tau^2/m I_m\big)^{-1}\Sigma^{1/2}\boldsymbol{\delta}\big\|^2\big\} \\
&= E\big[\operatorname{tr}\big\{\tau^2/m\,\Sigma^2\big(\Sigma + \tau^2/m I_m\big)^{-2}\big\}\big] \\
&\quad + E\big[\operatorname{tr}\big\{\big(\tau^2/m\big)^2\Sigma\big(\Sigma + \tau^2/m I_m\big)^{-2}\big\}\big] \\
&= E\big[\operatorname{tr}\big\{\tau^2/m\,\Sigma\big(\Sigma + \tau^2/m I_m\big)^{-1}\big\}\big] \\
&= E\big[\operatorname{tr}\big\{\big(\Sigma^{-1} + m/\tau^2 I_m\big)^{-1}\big\}\big].
\end{aligned} \tag{32}$$

Thus, Brown's identity and (32) imply

$$\begin{aligned}
R_{\text{seq}}\big\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\big\} - R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} &= E\big[\operatorname{tr}\big\{\Sigma^2 I_\Sigma\big(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}\big)\big\}\big] \\
&\quad + E\big[\operatorname{tr}\big\{\big(\Sigma^{-1} + m/\tau^2 I_m\big)^{-1}\big\}\big] - E\big\{\operatorname{tr}(\Sigma)\big\} \\
&= E\big[\operatorname{tr}\big\{\Sigma^2 I_\Sigma\big(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}\big)\big\}\big] \\
&\quad - E\big[\operatorname{tr}\big\{\Sigma^2\big(\Sigma + \tau^2/m I_m\big)^{-1}\big\}\big].
\end{aligned}$$

Taking $\mathbf{u} = \boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1$, $\mathbf{v} = (\Sigma - \gamma\sigma_m I_m)^{1/2}\boldsymbol{\delta}_2$, and $\Psi = \Sigma^2$ in Stam's inequality, where $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, and $0 < \gamma < 1$ are given in (31), one obtains

$$\begin{aligned} &R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} \\ &\quad\le E\{\mathrm{tr}(\Sigma^2[I_\Sigma\{\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1\}^{-1} + \Sigma - \gamma\sigma_m I_m]^{-1})\} \\ &\qquad - E[\mathrm{tr}\{\Sigma^2(\Sigma + \tau^2/m I_m)^{-1}\}]. \end{aligned}$$

By orthogonal invariance, $I_\Sigma\{\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1\} = \zeta I_m$ for some $\zeta \ge 0$. Thus,

$$\begin{aligned} R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} &\le E\left(\mathrm{tr}\left[\Sigma^2\left\{\Sigma + \left(\frac{1}{\zeta} - \gamma\sigma_m\right)I_m\right\}^{-1}\right]\right) \\ &\quad - E[\mathrm{tr}\{\Sigma^2(\Sigma + \tau^2/m I_m)^{-1}\}]. \end{aligned} \tag{33}$$

Next we bound ζ. Conditioning on Σ, applying Brown's identity with $\gamma\sigma_m I_m$ in place of Σ, and applying Marchand's inequality (Lemma B.3) with $\nu^2 = \gamma\sigma_m$, we obtain

$$m\gamma^2\sigma_m^2\zeta = \mathrm{tr}[\gamma^2\sigma_m^2 I_\Sigma\{\boldsymbol{\theta} + (\gamma\sigma_m)^{1/2}\boldsymbol{\delta}_1\}] \le m\gamma\sigma_m - \left(1 - \frac{1}{m}\right)\frac{\tau^2\gamma\sigma_m m}{\tau^2 + \gamma\sigma_m m}.$$

Dividing by $m\gamma^2\sigma_m^2$ above, it follows that

$$\zeta \le \left(\frac{1}{\gamma\sigma_m}\right)\frac{\gamma\sigma_m + \tau^2/m^2}{\gamma\sigma_m + \tau^2/m}.$$

Further rearranging implies that

$$\frac{1}{\zeta} - \gamma\sigma_m \ge (m-1)\frac{\gamma\sigma_m\tau^2}{\gamma\sigma_m m^2 + \tau^2}.$$

Hence, combining this with (33),

$$\begin{aligned} R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} &\le E\left(\mathrm{tr}\left[\Sigma^2\left\{\Sigma + (m-1)\frac{\gamma\sigma_m\tau^2}{\gamma\sigma_m m^2 + \tau^2}I_m\right\}^{-1}\right]\right) \\ &\quad - E[\mathrm{tr}\{\Sigma^2(\Sigma + \tau^2/m I_m)^{-1}\}]. \end{aligned}$$

Finally, taking $\gamma \uparrow 1$ above yields

$$\begin{aligned} R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_r(\tau), \boldsymbol{\theta}\} - R_{\mathrm{seq}}\{\hat{\boldsymbol{\theta}}_{S^{m-1}(\tau)}, \boldsymbol{\theta}\} &\le E\left(\mathrm{tr}\left[\Sigma^2\left\{\Sigma + (m-1)\frac{\sigma_m\tau^2}{\sigma_m m^2 + \tau^2}I_m\right\}^{-1}\right]\right) \\ &\quad - E[\mathrm{tr}\{\Sigma^2(\Sigma + \tau^2/m I_m)^{-1}\}] \\ &\le E\left[\left(\frac{\sigma_1}{m\sigma_m} \wedge 1\right)\mathrm{tr}\{(\Sigma^{-1} + m/\tau^2 I_m)^{-1}\}\right], \end{aligned}$$

where it is elementary to verify the second inequality upon diagonalizing Σ. This completes the proof of the lemma. □

Theorem 2(a) follows immediately from (28) and Lemmas B.2 and B.6.

B.4. Proof of Theorem 2(b)

It remains to prove Theorem 2(b), which is achieved through a sequence of lemmas. Similar to the proof of Theorem 2(a), the initial steps involve reducing the problem from the linear model to the sequence model. In the following lemma, we derive a basic property of orthogonally equivariant estimators for $\boldsymbol{\beta}$ (in the linear model) when $d > n$.

Lemma B.7. *Suppose $d > n$ and that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X) \in \mathscr{E}$ is an orthogonally equivariant estimator for $\boldsymbol{\beta}$ in the linear model (1). Further suppose that $X = UDV^T$, where $U \in \mathcal{O}(n)$, D is an $n \times n$ diagonal matrix, and V is an $n \times d$ matrix with orthonormal columns. Let V_0 be a $(d-n) \times d$ matrix so that $(V \;\; V_0) \in \mathcal{O}(d)$. Then $V_0^T \hat{\boldsymbol{\beta}} = 0$.*

Proof. Let $W \in \mathcal{O}(d-n)$ and let $V_W = VV^T + V_0 W V_0^T \in \mathcal{O}(d)$. Then

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = V_W \hat{\boldsymbol{\beta}}(\mathbf{y}, XV_W) = V_W \hat{\boldsymbol{\beta}}(\mathbf{y}, X). \tag{34}$$

Since (34) holds for all $W \in \mathcal{O}(d-n)$, we must have $V_0^T \hat{\boldsymbol{\beta}} = 0$. □

In the next lemma, we relate the minimax risk under the linear model $r(\tau)$ to the risk under the sequence model.

Lemma B.8. *Suppose that $d > n$ and let $\tau^2 > 0$. In the sequence model (26), suppose that $m = n$ and $\Sigma = (XX^T)^{-1}$. For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$, let $\boldsymbol{\theta}_n = (\theta_1, \ldots, \theta_n)^T \in \mathbb{R}^n$ be the projection onto the first n coordinates. Then*

$$r(\tau) \geq \int_{S^{d-1}(\tau)} r_{\mathrm{seq}}\big(\|\boldsymbol{\theta}_n\|\big)\, \mathrm{d}\pi_{S^{n-1}(\tau)}(\boldsymbol{\theta}) + \frac{d-n}{n}\tau^2.$$

Proof. By Proposition 3,

$$r(\tau) = \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} \int_{S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}). \tag{35}$$

Assume that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X) \in \mathscr{E}$ and let $X = UDV^T$ be the decomposition in Lemma B.7. Additionally, let $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_1, \ldots, \hat{\beta}_n)^T \in \mathbb{R}^n$ be the first n coordinates of $\hat{\boldsymbol{\beta}}$. Then, under the linear model (1),

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 &= \big\|V^T\hat{\boldsymbol{\beta}}(\mathbf{y}, X) - V^T\boldsymbol{\beta}\big\|^2 + \big\|V_0^T\boldsymbol{\beta}\big\|^2 \\ &= \big\|\hat{\boldsymbol{\beta}}_n\big\{UDV^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}, (UD \;\; 0)\big\} - V^T\boldsymbol{\beta}\big\|^2 + \big\|V_0^T\boldsymbol{\beta}\big\|^2.\end{aligned}$$

Let $\boldsymbol{\beta}_n = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n$. Integrating $\boldsymbol{\beta}$ over $S^{d-1}(\tau)$ with respect to the uniform measure, making the change of variables

$$\boldsymbol{\beta} \mapsto (V \;\; V_0) \begin{pmatrix} U^T & 0 \\ 0 & I_{d-n} \end{pmatrix} \boldsymbol{\beta},$$

and using the fact that $\hat{\boldsymbol{\beta}} \in \mathscr{E}$, it follows that

$$\begin{aligned}
&\int_{S^{d-1}(\tau)} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) \\
&\quad = \int_{S^{d-1}(\tau)} \left\| \hat{\boldsymbol{\beta}}_n\{UDV^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}, (UD \;\; 0)\} - V^T\boldsymbol{\beta} \right\|^2 \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) + \frac{d-n}{d}\tau^2 \\
&\quad = \int_{S^{d-1}(\tau)} \left\| \hat{\boldsymbol{\beta}}_n\{UDU^T\boldsymbol{\beta}_n + \boldsymbol{\varepsilon}, (UD \;\; 0)\} - U^T\boldsymbol{\beta}_n \right\|^2 \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) + \frac{d-n}{d}\tau^2 \\
&\quad = \int_{S^{d-1}(\tau)} \left\| \hat{\boldsymbol{\beta}}_n\{UDU^T\boldsymbol{\beta}_n + \boldsymbol{\varepsilon}, \left(UDU^T \;\; 0\right)\} - \boldsymbol{\beta}_n \right\|^2 \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) + \frac{d-n}{d}\tau^2.
\end{aligned}$$

Next, for $\mathbf{w} \in \mathbb{R}^n$ and $n \times n$ positive definite matrices Σ, define the estimator for the sequence model $\hat{\boldsymbol{\theta}}(\mathbf{w}, \Sigma) = \hat{\boldsymbol{\beta}}_n\{\Sigma^{-1/2}\mathbf{w}, (\Sigma^{-1/2} \;\; 0)\}$. Then, with $m = n$ and $\Sigma = (XX^T)^{-1} = UD^{-2}U^T$,

$$\begin{aligned}
&\int_{S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) \\
&\quad = \int_{S^{d-1}(\tau)} R_{\mathrm{seq}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_n) \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\theta}) + \frac{d-n}{n}\tau^2.
\end{aligned}$$

By equivariance, $R_{\mathrm{seq}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})$ is constant over spheres $\boldsymbol{\vartheta} \in S^{d-1}(\|\boldsymbol{\theta}_n\|)$, which implies that $R_{\mathrm{seq}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_n) \geq r_{\mathrm{seq}}(\|\boldsymbol{\theta}_n\|)$. Hence,

$$\int_{S^{d-1}(\tau)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\beta}) \geq \int_{S^{d-1}(\tau)} r_{\mathrm{seq}}\left(\|\boldsymbol{\theta}_n\|\right) \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\theta}) + \frac{d-n}{n}\tau^2. \tag{36}$$

The lemma follows from (35) and (36). □

The proof of Theorem 2(b) will follow from a calculation involving Lemmas B.6 and B.8. The key part of this calculation is contained in the following lemma.

Lemma B.9. *Suppose that* $2 < n < d$. *Let* $s_1 \geq \cdots \geq s_n \geq 0$ *denote the nonzero eigenvalues of* $n^{-1}X^TX$. *Then*

$$r(\tau) \geq E\left[\left(1 - \frac{s_1}{ns_n}\right) \mathrm{tr}\left\{\left(XX^T + \frac{n(d-2)}{\tau^2(n-2)} I_n\right)^{-1}\right\}\right] + \frac{d-n}{d}\tau^2.$$

Proof. With $\boldsymbol{\theta}_n \in \mathbb{R}^n$, $m = n$ and $\Sigma = (XX^T)^{-1}$, Lemma B.6 and (32) imply that

$$\begin{aligned} r_{\text{seq}}\big(\|\boldsymbol{\theta}_n\|\big) &= R_{\text{seq}}\{\hat{\boldsymbol{\theta}}_{S^{n-1}(\|\boldsymbol{\theta}_n\|)}, \boldsymbol{\theta}_n\} \\ &\geq R_{\text{seq}}\big\{\hat{\boldsymbol{\theta}}_r\big(\|\boldsymbol{\theta}_n\|\big), \boldsymbol{\theta}_n\big\} - E\bigg[\bigg(\frac{s_1}{ns_n} \wedge 1\bigg)\operatorname{tr}\bigg\{\bigg(XX^T + \frac{n}{\|\boldsymbol{\theta}_n\|^2} I_n\bigg)^{-1}\bigg\}\bigg] \qquad (37) \\ &= E\bigg[\bigg\{\bigg(1 - \frac{s_1}{ns_n}\bigg) \vee 0\bigg\}\operatorname{tr}\bigg\{\bigg(XX^T + \frac{n}{\|\boldsymbol{\theta}_n\|^2} I_n\bigg)^{-1}\bigg\}\bigg]. \end{aligned}$$

Additionally, if $\boldsymbol{\theta} \sim \pi_{S^{d-1}(\tau)}$, then $\boldsymbol{\theta} = \tau \mathbf{z}/\|\mathbf{z}\|$ in distribution, where $\mathbf{z} \sim N(0, I_d)$; using basic properties of the chi-squared distribution, it follows that

$$\int_{S^{d-1}(\tau)} \frac{1}{\|\boldsymbol{\theta}_n\|^2} \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\theta}) = \frac{d-2}{\tau^2(n-2)}, \qquad (38)$$

where $\boldsymbol{\theta}_n = (\theta_1, \ldots, \theta_n)^T \in \mathbb{R}^n$ is the projection of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$ onto the first n coordinates. Thus, by (37), Jensen's inequality and (38),

$$\begin{aligned} &\int_{S^{d-1}d(\tau)} r\big(\|\boldsymbol{\theta}_n\|\big) \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\theta}) \\ &\geq E\bigg[\bigg(1 - \frac{s_1}{ns_n}\bigg)\operatorname{tr}\bigg\{\bigg(XX^T + \int_{S^{d-1}(\tau)} \frac{n}{\|\boldsymbol{\theta}_n\|^2} \, \mathrm{d}\pi_{S^{d-1}(\tau)}(\boldsymbol{\theta}) I_n\bigg)^{-1}\bigg\}\bigg] \\ &\geq E\bigg[\bigg(1 - \frac{s_1}{ns_n}\bigg)\operatorname{tr}\bigg\{\bigg(XX^T + \frac{n(d-2)}{\tau^2(n-2)} I_n\bigg)^{-1}\bigg\}\bigg]. \end{aligned}$$

The lemma follows by combining the last inequality above with Lemma B.8. □

We now have the tools to complete the proof of Theorem 2(b). Suppose that $d > n$ and $\boldsymbol{\beta} \in S^{d-1}(\tau)$. Then

$$R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} = E\big[\operatorname{tr}\big\{\big(XX^T + d/\tau^2 I_n\big)^{-1}\big\}\big] + \frac{d-n}{d}\tau^2.$$

Since $R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} = r(\tau)$, Lemma B.9 implies

$$\begin{aligned} R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} - R\{\hat{\boldsymbol{\beta}}_{S^{d-1}(\tau)}, \boldsymbol{\beta}\} &= R\big\{\hat{\boldsymbol{\beta}}_r(\tau), \boldsymbol{\beta}\big\} - r(\tau) \\ &\leq E\big[\operatorname{tr}\big\{\big(XX^T + d/\tau^2 I_n\big)^{-1}\big\}\big] \\ &\quad - E\bigg[\bigg(1 - \frac{s_1}{ns_n}\bigg)\operatorname{tr}\bigg\{\bigg(XX^T + \frac{n(d-2)}{\tau^2(n-2)} I_n\bigg)^{-1}\bigg\}\bigg] \\ &\leq \frac{1}{n} E\bigg[\frac{s_1}{s_n}\operatorname{tr}\bigg\{\bigg(XX^T + \frac{d}{\tau^2} I_n\bigg)^{-1}\bigg\}\bigg] \\ &\quad + \frac{2(d-n)}{\tau^2(n-2)} E\bigg[\operatorname{tr}\bigg\{\bigg(XX^T + \frac{d}{\tau^2} I_n\bigg)^{-2}\bigg\}\bigg]. \end{aligned}$$

Theorem 2(b) follows.

Appendix C

Lemma C.1. *Let $s_d \geq 0$ denote the smallest eigenvalue of $n^{-1}X^TX$. Suppose that $a > 0$ is a positive real number and that $n - d \geq 2a + 1$. If $d = 1$, then $E(s_d^{-a}) \leq \mathrm{e}^a$. If $d \geq 2$, then*

$$E\bigl(s_d^{-a}\bigr) \leq 2\left\{\frac{\pi}{4}\sqrt{\frac{n^5}{(d-1)(n-d)^2}}\mathrm{e}^{n+1/2}\right\}^{2a/(n-d+1)}. \tag{39}$$

Proof. Suppose first that $d = 1$. Then $ns_d \sim \chi_n^2$ is a chi-squared random variable on n degrees of freedom. By Theorem 1 of Kečkić and Vasić [36], which gives convenient bounds on the ratio of two gamma functions,

$$E\bigl(s_d^{-a}\bigr) = \frac{(n/2)^a\Gamma(n/2-a)}{\Gamma(n/2)} \leq \frac{(n/2)^a(n/2-a)^{n/2-a-1}}{(n/2)^{n/2-1}}\mathrm{e}^a \leq \mathrm{e}^a.$$

This proves the first part of the lemma.

Now suppose that $d \geq 2$. Suppose further that (39) is true for $a = 1$. If $0 < a_0 < 1$, then

$$E\bigl(s_d^{-a_0}\bigr) \leq \bigl\{E\bigl(s_d^{-1}\bigr)\bigr\}^{a_0} \leq 2\left\{\frac{\pi}{4}\sqrt{\frac{n^5}{(d-1)(n-d)^2}}\mathrm{e}^{n+1/2}\right\}^{2a_0/(n-d+1)}$$

and (39) holds for $a = a_0$. Thus, we may assume that $a \geq 1$. Let $t > 0$ be a fixed positive number. Then

$$E\bigl(s_d^{-a}\bigr) \leq E\bigl[s_d^{-a}\mathbf{1}_{\{s_d \leq t\}}\bigr] + t^{-a}. \tag{40}$$

Muirhead [41] (Corollary 3.2.19) gives the joint density of the ordered eigenvalues, $s_1 > \cdots > s_d > 0$, of $n^{-1}X^TX$:

$$f_{d,n}(s_1,\ldots,s_d) = c_{d,n}\exp\left(-\frac{n}{2}\sum_{j=1}^d s_j\right)\prod_{j=1}^d s_j^{(n-d-1)/2}\prod_{i<j}(s_i - s_j),$$

where

$$c_{d,n} = \frac{\pi^{d^2/2}}{(2/n)^{dn/2}\Gamma_d(d/2)\Gamma_d(n/2)}$$

and

$$\Gamma_d(n/2) = \pi^{d(d-1)/4}\prod_{j=1}^d \Gamma\bigl\{(n-j+1)/2\bigr\}$$

is the multivariate gamma function. Let $T_d = \{(s_1, \ldots, s_d) \in \mathbb{R}^d; s_1 > \cdots > s_d > 0\}$. Then,

$$\begin{aligned}
E\big[s_d^{-a}\mathbf{1}_{\{s_d<t\}}\big] &= \int_{T_d\cap\{s_d<t\}} s_d^{-a} f_{d,n}(s_1, \ldots, s_d)\,\mathrm{d}s_1\cdots\mathrm{d}s_d \\
&\le \int_{T_{d-1}} \left\{\int_0^t s_d^{-a} f_{d,n}(s_1, \ldots, s_d)\,\mathrm{d}s_d\right\}\mathrm{d}s_1\cdots\mathrm{d}s_{d-1} \\
&\le \frac{c_{d,n}}{c_{d-1,n}} \int_{T_{d-1}} \left(\prod_{j=1}^{d-1} s_j\right)^{1/2} f_{d-1,n}(s_1, \ldots, s_{d-1})\,\mathrm{d}s_1\cdots\mathrm{d}s_{d-1} \\
&\quad\cdot \int_0^t s^{(n-d-1)/2-a}\mathrm{e}^{-ns/2}\,\mathrm{d}s \\
&\le \frac{c_{d,n}}{c_{d-1,n}} E\big\{\det\big(n^{-1}Z^TZ\big)^{1/2}\big\}t^{(n-d+1)/2-a},
\end{aligned}$$

where Z is an $n\times(d-1)$-dimensional matrix with i.i.d. $N(0,1)$ entries and the last inequality above follows from the fact that $n-d\ge 2a+1$. It is easy to check that

$$\frac{c_{n,d}}{c_{n,d-1}} = \frac{\sqrt{\pi}(n/2)^{n/2}}{\Gamma(d/2)\Gamma\{(n-d+1)/2\}}.$$

Additionally, it is well known (Problem 3.11 in Muirhead [41], for instance) that

$$E\big\{\det\big(n^{-1}Z^TZ\big)^{1/2}\big\} = (2/n)^{(d-1)/2}\frac{\Gamma\{(n+1)/2\}}{\Gamma\{(n-d+1)/2\}}.$$

By Corollary 1.2 of Batir [4] (a variant of Stirling's approximation),

$$x^x\mathrm{e}^{-x}\sqrt{2x+1} \le \Gamma(x+1) \le x^x\mathrm{e}^{-x}\sqrt{\pi(2x+1)}, \qquad \text{for all } x\ge 0.$$

It follows that,

$$\begin{aligned}
\frac{c_{n,d}}{c_{n,d-1}} E\big\{\det\big(n^{-1}Z^TZ\big)^{1/2}\big\} &= \frac{\sqrt{\pi}(n/2)^{(n-d+1)/2}\Gamma\{(n+1)/2\}}{\Gamma(d/2)\Gamma\{(n-d+1)/2\}^2} \\
&\le \frac{\pi n^{(n-d+2)/2}(n-1)^{(n-1)/2}\mathrm{e}^{(n-d-3)/2}}{4(d-2)^{(d-2)/2}\sqrt{d-1}(n-d-1)^{n-d-1}(n-d)} \\
&\le \frac{\pi}{4}\sqrt{\frac{n^5}{(d-1)(n-d)^2}}\mathrm{e}^{n+1/2}
\end{aligned}$$

and

$$E\big[s_d^{-a}\mathbf{1}_{\{s_d<t\}}\big] \le t^{(n-d+1)/2-a}\frac{\pi}{4}\sqrt{\frac{n^5}{(d-1)(n-d)^2}}\mathrm{e}^{n+1/2}.$$

Thus, by (40)

$$E\big(s_d^{-a}\big) \le t^{(n-d+1)/2-a}\frac{\pi}{4}\sqrt{\frac{n^5}{(d-1)(n-d)^2}}\mathrm{e}^{n+1/2} + t^{-a}.$$

Taking $t = [(\pi/4)\sqrt{n^5/\{(d-1)(n-d)^2\}}\mathrm{e}^{n+1/2}]^{-2/(n-d+1)}$ gives (39). □

Lemma C.2. *Let $s_1 \ge s_d \ge 0$ denote the largest and smallest eigenvalues of $n^{-1}X^TX$, respectively. Suppose that $a > 0$ is a fixed positive real number and that $0 < d/n \le \rho_+ < 1$ for some fixed constant $\rho_+ \in \mathbb{R}$.*

(a) $E(s_1^a) = \mathrm{O}(1)$.
(b) *If $n - d > 2a + 1$, then $E(s_d^{-a}) = \mathrm{O}(1)$.*

The constants implicit in the bounds from parts (a) *and* (b) *depend on the exponent a.*

Proof. Part (a) is well known and may be easily derived from large deviations results for s_1 (see, e.g., Theorem II.13 of Davidson and Szarek [22]). Part (b) follows directly from Lemma C.1. □

Lemma C.3. *Let $a > 0$ be a fixed positive real number. If $n > 2a$, then*

$$\sup_{\boldsymbol{\beta}\in S^{d-1}(\tau)} E_{\boldsymbol{\beta}}\left\{\left(\frac{1}{\hat\tau^2 + d/n}\right)^a\right\} = \mathrm{O}\left\{\left(\frac{n/d+1}{\tau^2+1}\right)^a\right\},$$

*where the implicit constant in the big-*O *bound depends on the exponent a.*

Proof. Suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$. Since $\|\mathbf{y}\|^2/(\tau^2+1) \sim \chi_n^2$ has a chi-squared distribution with n degrees of freedom,

$$\begin{aligned} E_{\boldsymbol{\beta}}\left\{\left(\frac{1}{\hat\tau^2 + d/n}\right)^a\right\} &\le E_{\boldsymbol{\beta}}\left\{\left(\frac{n/d+1}{\hat\tau^2+1}\right)^a\right\} \\ &\le (n/d+1)^a n^a E_{\boldsymbol{\beta}}\big(\|\mathbf{y}\|^{-2a}\big) \\ &= \mathrm{O}\left\{\left(\frac{n/d+1}{\tau^2+1}\right)^a\right\}. \end{aligned}$$

□

Lemma C.4. *Let $P_{\boldsymbol{\beta}}(\cdot)$ denote the probability measure induced by the joint distribution of $(\mathbf{y}, X)$, where $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Then*

$$\sup_{\boldsymbol{\beta}\in S^{d-1}(\tau)} P_{\boldsymbol{\beta}}\big(\hat\tau^2 = 0\big) \le \mathrm{e}^{(-n/4)(\tau^2/(\tau^2+1))^2}.$$

Proof. Suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$. Let $t \ge 0$ be fixed. Since $V = \|\mathbf{y}\|^2/(\tau^2+1) \sim \chi_n^2$ has a chi-squared distribution with n degrees of freedom, it follows that

$$P_{\boldsymbol{\beta}}\big(\hat\tau^2 = 0\big) = P_{\boldsymbol{\beta}}\left(V \le \frac{n}{\tau^2+1}\right) \le \mathrm{e}^{nt/(\tau^2+1)}E_{\boldsymbol{\beta}}\big(\mathrm{e}^{-tV}\big) = \left(\frac{\mathrm{e}^{2t/(\tau^2+1)}}{1+2t}\right)^{n/2}.$$

Taking $t = \tau^2/2$ and using the fact that $(1-x)\mathrm{e}^x \leq \mathrm{e}^{-x^2/2}$ for all $x \geq 0$ yields

$$P_{\boldsymbol{\beta}}\big(\hat{\tau}^2 = 0\big) \leq \left(\frac{\mathrm{e}^{\tau^2/(\tau^2+1)}}{\tau^2+1}\right)^{n/2} \leq \mathrm{e}^{(-n/4)(\tau^2/(\tau^2+1))^2}. \qquad \square$$

Lemma C.5. *Suppose $a > 0$ is a fixed positive real number. Then*

$$\sup_{\boldsymbol{\beta} \in S^{d-1}(\tau)} E_{\boldsymbol{\beta}}\big(\big|\hat{\tau}^2 - \tau^2\big|^a\big) = \mathrm{O}\left(\frac{\tau^{2a}+1}{n^{a/2}}\right),$$

*where the implicit constant in the big-*O *bound depends on the exponent a.*

Proof. Suppose that $\boldsymbol{\beta} \in S^{d-1}(\tau)$. From the definition of $\hat{\tau}^2$,

$$E_{\boldsymbol{\beta}}\big(\big|\hat{\tau}^2 - \tau^2\big|^a\big) \leq E_{\boldsymbol{\beta}}\left\{\left|\frac{1}{n}\|\mathbf{y}\|^2 - \big(\tau^2+1\big)\right|^a\right\} + \tau^{2a} P_{\boldsymbol{\beta}}\big(\hat{\tau}^2 = 0\big). \tag{41}$$

Since $\|\mathbf{y}\|^2/(\tau^2+1) \sim \chi^2_n$,

$$E_{\boldsymbol{\beta}}\left\{\left|\frac{1}{n}\|\mathbf{y}\|^2 - \big(\tau^2+1\big)\right|^a\right\} = \mathrm{O}\left(\frac{\tau^{2a}+1}{n^{a/2}}\right). \tag{42}$$

Additionally, Lemma C.4 implies

$$\begin{aligned}
\tau^{2a} P_{\boldsymbol{\beta}}\big(\hat{\tau}^2 = 0\big) &\leq \tau^{2a}\mathrm{e}^{(-n/4)(\tau^2/(\tau^2+1))^2} \\
&= \big(\tau^2+1\big)^a \left(\frac{\tau^2}{\tau^2+1}\right)^a \mathrm{e}^{(-n/4)(\tau^2/(\tau^2+1))^2} \\
&\leq \big(\tau^2+1\big)^a \left(\frac{2a}{n}\right)^{a/2} \mathrm{e}^{-a/2}.
\end{aligned} \tag{43}$$

The lemma follows by combining (41) and (43). $\square$

Acknowledgements

The author would like to thank Alan Edelman, Bill Strawderman, Cun-Hui Zhang and Sihai Zhao for their helpful comments and inspiration. The author thanks the Associate Editor and the referees for their careful reading of the paper and their suggestions that helped to greatly improve its presentation. Supported by NSF Grant DMS-1208785.

References

[1] Bai, Z.D. (1993). Convergence rate of expected spectral distributions of large random matrices. II. Sample covariance matrices. *Ann. Probab.* **21** 649–672. MR1217560

[2] Bai, Z.D., Miao, B. and Yao, J.-F. (2003). Convergence rates of spectral distributions of large sample covariance matrices. *SIAM J. Matrix Anal. Appl.* **25** 105–127. MR2002902
[3] Baranchik, A.J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.* **1** 312–321. MR0348928
[4] Batir, N. (2008). Inequalities for the gamma function. *Arch. Math. (Basel)* **91** 554–563. MR2465874
[5] Belitser, E.N. and Levit, B.Y. (1995). On minimax filtering over ellipsoids. *Math. Methods Statist.* **4** 259–273. MR1355248
[6] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324
[7] Beran, R. (1996). Stein estimation in high dimensions: A retrospective. In *Research Developments in Probability and Statistics* 91–110. Utrecht: VSP. MR1462411
[8] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR0804611
[9] Bickel, P.J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309. MR0630112
[10] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469
[11] Bondar, J.V. and Milnes, P. (1981). Amenability: A survey for statistical applications of Hunt–Stein and related conditions on groups. *Z. Wahrsch. Verw. Gebiete* **57** 103–128. MR0623457
[12] Borel, É. (1914). *Introduction Géométrique à Quelques Théories Physiques*. Paris: Gauthier-Villars.
[13] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136. MR0696857
[14] Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903. MR0286209
[15] Brown, L.D. (1990). The 1985 Wald Memorial Lectures. An ancillarity paradox which appears in multiple linear regression. *Ann. Statist.* **18** 471–538. With discussion and a reply by the author. MR1056325
[16] Brown, L.D. and Gajek, L. (1990). Information inequalities for the Bayes risk. *Ann. Statist.* **18** 1578–1594. MR1074424
[17] Brown, L.D. and Low, M.G. (1991). Information inequality bounds on the minimax risk (with an application to nonparametric regression). *Ann. Statist.* **19** 329–337. MR1091854
[18] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149
[19] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644
[20] Dalalyan, A. and Chen, Y. (2012). Fused sparsity and robust estimation for linear models with unknown variance. *Adv. Neural Inf. Process. Syst.* **25** 1268–1276.
[21] DasGupta, A. (2010). False vs. missed discoveries, Gaussian decision theory, and the Donsker–Varadhan principle. In *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown. Inst. Math. Stat. Collect.* **6** 1–21. Beachwood, OH: IMS. MR2798507
[22] Davidson, K.R. and Szarek, S.J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces* **I** 317–366. Amsterdam: North-Holland. MR1863696
[23] Diaconis, P. and Freedman, D. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré Probab. Stat.* **23** 397–423. MR0898502
[24] Dicker, L.H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat.* **7** 1806–1834. MR3084672
[25] Dicker, L.H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284.

[26] Donoho, D.L. and Johnstone, I.M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886
[27] Draper, N.R. and Van Nostrand, R.C. (1979). Ridge regression and James–Stein estimation: Review and comments. *Technometrics* **21** 451–466. MR0555086
[28] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012
[29] El Karoui, N. and Kösters, H. (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. Preprint. Available at arXiv:1105.1404.
[30] Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 37–65. MR2885839
[31] Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.* **29** 1601–1619. MR1891740
[32] Goldenshluger, A. and Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *J. Multivariate Anal.* **84** 40–60. MR1965822
[33] Golubev, G.K. (1987). Adaptive asymptotically minimax estimates for smooth signals. *Probl. Inf. Transm.* **23** 57–67. MR0893970
[34] Golubev, G.K. (1990). Quasilinear estimates for signals in L_2. *Probl. Inf. Transm.* **26** 15–20. MR1051584
[35] Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
[36] Kečkić, J.D. and Vasić, P.M. (1971). Some inequalities for the gamma function. *Publ. Inst. Math. (Beograd) (N.S.)* **11** 107–114. MR0308446
[37] Leeb, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.* **37** 2838–2876. MR2541449
[38] Lévy, P. (1922). *Leçons d'Analyse Fonctionnelle*. Paris: Gauthier-Villars.
[39] Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR–Sb.* **1** 457–483.
[40] Marchand, E. (1993). Estimation of a multivariate mean with constraints on the norm. *Canad. J. Statist.* **21** 359–366. MR1254283
[41] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0652932
[42] Nussbaum, M. (1999). Minimax risk: Pinsker bound. In *Encyclopedia of Statistical Sciences Update Vol.* 3 451–460. New York: Wiley.
[43] Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 52–68. MR0624591
[44] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274
[45] Robert, C. (1990). Modified Bessel functions and their applications in probability and statistics. *Statist. Probab. Lett.* **9** 155–161. MR1045178
[46] Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408
[47] Stam, A.J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. Control* **2** 101–112. MR0109101
[48] Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics* 424–443. Stanford, CA: Stanford Univ. Press. MR0120718
[49] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[50] Tihonov, A.N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR* **151** 501–504. MR0162377

[51] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359

[52] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

[53] Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inform. Theory* **44** 1246–1250. MR1616672

[54] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

Received January 2013 and revised September 2013