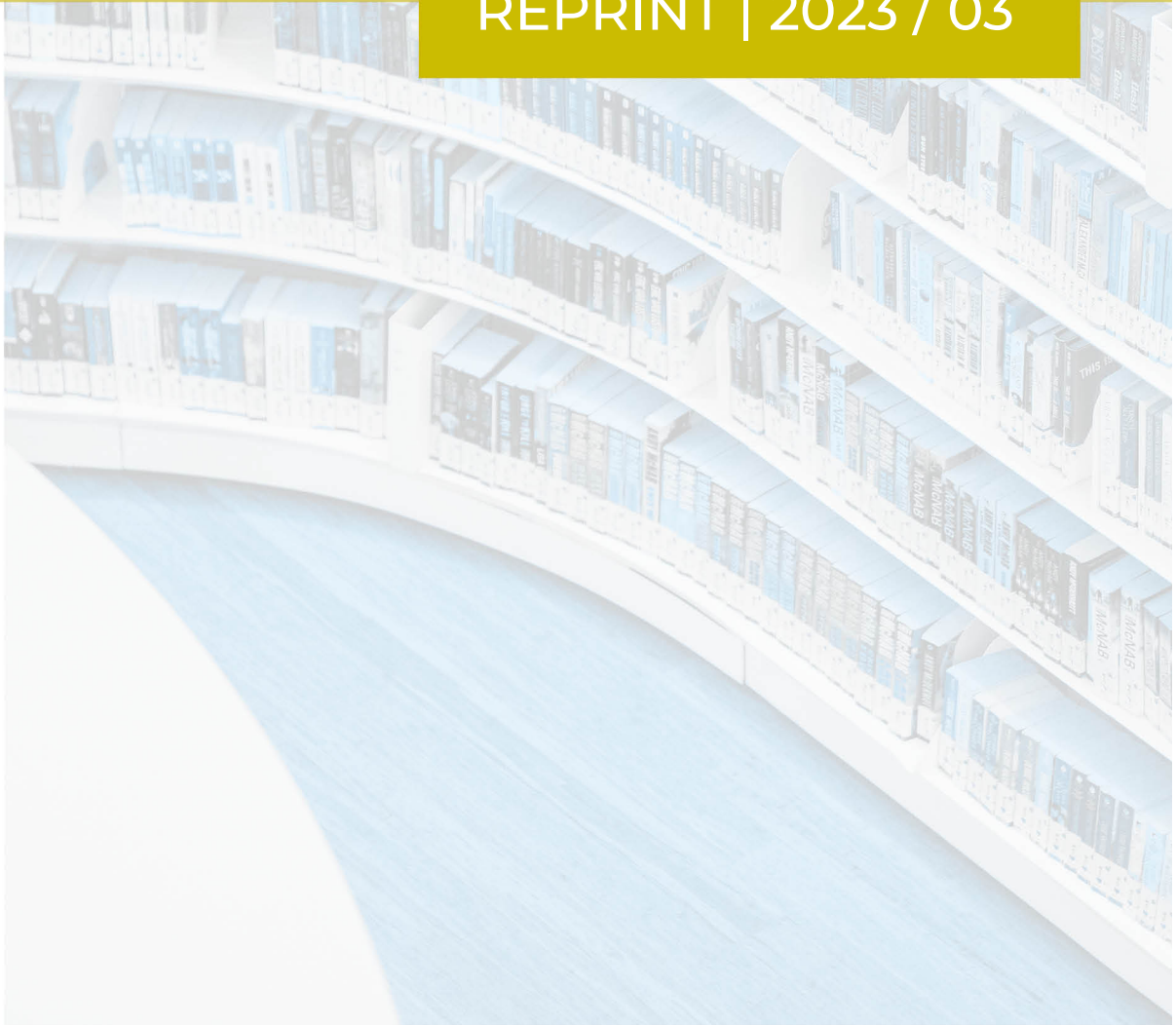


AN ANALYTICAL SHRINKAGE ESTIMATOR FOR LINEAR REGRESSION

Nathan Lassance

REPRINT | 2023 / 03



LFIN

Voie du Roman Pays 34, L1.03.01 B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/lfin/publications.html>

An Analytical Shrinkage Estimator for Linear Regression

Nathan Lassance^a

^aLFIN/LIDAM, UCLouvain, Chaussée de Binche 151, Mons, 5000, Belgium, nathan.lassance@uclouvain.be

Abstract

We derive an analytical solution to the optimal shrinkage of OLS regression coefficients toward a constant target, under any first two moments of predictors. The estimator closely mimics the prediction performance of ridge penalty, which admits no general analytical solution.

Keywords: linear regression, prediction error, shrinkage, out-of-sample

1. Introduction

The ordinary least squares (OLS) estimator for linear regression offers notoriously poor prediction out of sample due to its large estimation errors. The most popular technique to improve upon OLS is to *shrink* OLS coefficients by adding penalties to the objective function. Examples are ridge [3], lasso [13], elastic net [14], and principal component (PC) lasso [12]. The common challenge in these methods is the calibration of the penalty coefficients, for which several approaches exist such as cross-validation, error bounds, information criteria, and adaptive validation; see [6, Chapter 4] for an overview. However, no analytical solution exists in general for the penalty coefficients delivering the best out-of-sample prediction, and for this reason they are deemed “difficult to calibrate in practice” [7].

We explore an alternative approach in this paper, which is analytically solvable. Instead of adding penalties to the objective function, we directly shrink OLS coefficients toward a constant target. This target can be set arbitrarily or optimized; see part 2 of Proposition 3. We identify analytically the optimal shrinkage intensity that minimizes the expected out-of-sample prediction error, for which we provide a consistent estimator. This analytical solution is available in a general setting with any first two moments of predictors, which we achieve by deriving a closed-form expression for the covariance matrix of OLS coefficients in this general setting. The generality of our analytical solution contrasts with the penalty approach, for which analytical solutions exist only for restrictive distributional assumptions. For example, the optimal ridge penalty is known when predictors have zero mean and identity covariance matrix (i.e., are orthonormal) [8], but not in general.

We provide simulation results that give credence to the appeal of the proposed approach. We first compare our shrinkage estimator to the analytical ridge solution when predictors are orthonormal. Across different values of the sample size and dimension, our approach can closely mimic the performance of the ridge penalty. We then consider the general setting in which predictors are correlated and non-standardized, which our analytical approach can also accommodate. We find that

our shrinkage estimator delivers prediction errors close to those in the orthonormal setting despite the increased level of estimation errors coming from estimates of the mean and covariance matrix of predictors.

Moreover, we show that optimizing the constant target delivers lower prediction error than arbitrarily shrinking toward a zero vector as is implicit in the ridge penalty.

Other linear shrinkage estimators exist in the literature; see, e.g., the textbook of [1] and, more recently, [10]. As an example, in our simulations we compare our estimator with that of [9] that combines OLS and PC regression using a Stein-rule linear shrinkage estimator. We find that our approach offers a competitive performance. Moreover, unlike [9], our theory does not assume predictors are centered, and does not require calibrating hyperparameters such as the number of PCs to keep.

Shrinking OLS coefficients directly resembles the methodology introduced by [4] in the finance literature, which shrinks estimated portfolio weights to optimize the expected out-of-sample investment performance; see [5] for a review. Therefore, the present paper points to potentially fruitful applications of this literature to the linear regression problem.

For the sake of conciseness, we only give full proofs of theoretical results when they do not directly follow from previously exposed material. We report these proofs in Section 6.

2. The analytical shrinkage estimator

We consider the classical linear regression problem where the univariate response y is modeled as a linear function of p random predictors $\mathbf{x} = (x_1, \dots, x_p)^\top$,

$$y = \alpha + \boldsymbol{\beta}^\top \mathbf{x} + \epsilon, \quad (1)$$

where α is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ are the regression coefficients, the predictors $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the noise $\epsilon \sim N(0, \sigma^2)$, and $\mathbf{x} \perp \epsilon$. We denote the centered response by

$$\tilde{y} \equiv y - \alpha, \quad (2)$$

which is the random variable we aim to predict. We also denote

$$\theta^2 \equiv \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \quad (3)$$

Given an i.i.d. sample of $n \times p$ predictors $\mathbf{X}$ and $n \times 1$ noises $\boldsymbol{\epsilon}$ drawn from (1), the $n \times 1$ responses $\mathbf{y}$ follow the model

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (4)$$

where $\mathbf{1}_n$ is a vector of n ones, each column $\mathbf{x}_i$ of $\mathbf{X}$ is i.i.d. with $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. We assume that $n > p + 1$.

Our objective is to estimate $\boldsymbol{\beta}$ to minimize the expected *out-of-sample* prediction error. The classical OLS estimator of $\boldsymbol{\beta}$ is instead designed to minimize the *in-sample* prediction error:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ols} &\equiv \arg \min_{\boldsymbol{\beta}} \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \alpha \mathbf{1}_n) \\ &= \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}. \end{aligned} \quad (5)$$

Although $\hat{\boldsymbol{\beta}}_{ols}$ is an unbiased estimator of the true $\boldsymbol{\beta}$, it is notoriously deficient due to its large variance. Specifically, in the next proposition we derive a novel analytical expression for the covariance matrix of $\hat{\boldsymbol{\beta}}_{ols}$ in our general setting in which no assumption is made about the true and sample moments of $\mathbf{x}$.

Proposition 1 *The covariance matrix of $\hat{\boldsymbol{\beta}}_{ols}$ in (5) is*

$$\text{Var}[\hat{\boldsymbol{\beta}}_{ols}] = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] = \frac{\sigma^2}{n-p-1} (\boldsymbol{\Sigma}^{-1} - a \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}), \quad (6)$$

where

$$a = {}_1F_1\left(1; \frac{n+2}{2}; -\frac{n}{2}\theta^2\right), \quad (7)$$

${}_1F_1$ is the confluent hypergeometric function and θ^2 is defined in (3). Moreover, $a \rightarrow 1/(1+\theta^2)$ as $n \rightarrow \infty$.

We observe indeed that the elements in $\text{Var}[\hat{\boldsymbol{\beta}}_{ols}]$ are large if p is close to $n-1$, i.e., in high dimension. Note that $a \approx 1/(1+\theta^2)$ with very good accuracy even for reasonable values of n , which can be used to simplify some expressions below.

To improve upon OLS, we search for an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ having minimum *expected out-of-sample prediction error*,

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) \equiv \mathbb{E}[(\tilde{\mathbf{y}} - \hat{\boldsymbol{\beta}}^\top \mathbf{x})^2], \quad (8)$$

where the expectation is taken under the true model distribution (1). It can be shown that the prediction error $\mathcal{L}(\hat{\boldsymbol{\beta}})$ depends on the mean and covariance matrix of $\hat{\boldsymbol{\beta}}$ as follows:

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{\beta}}) &= \sigma^2 + (\mathbb{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{\mu}])^2 + \mathbb{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{\Sigma}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})] + \text{Var}[\hat{\boldsymbol{\beta}}^\top \boldsymbol{\mu}] \\ &= \sigma^2 + ((\boldsymbol{\beta} - \mathbb{E}[\hat{\boldsymbol{\beta}}])^\top \boldsymbol{\mu})^2 + (\boldsymbol{\beta} - \mathbb{E}[\hat{\boldsymbol{\beta}}])^\top \boldsymbol{\Sigma}(\boldsymbol{\beta} - \mathbb{E}[\hat{\boldsymbol{\beta}}]) \\ &\quad + \text{Trace}(\boldsymbol{\Sigma} \text{Var}[\hat{\boldsymbol{\beta}}]) + \boldsymbol{\mu}^\top \text{Var}[\hat{\boldsymbol{\beta}}] \boldsymbol{\mu}. \end{aligned} \quad (9)$$

Plugging $\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] = \boldsymbol{\beta}$ and $\text{Var}[\hat{\boldsymbol{\beta}}_{ols}]$ in (6) into (9), we obtain the prediction error of $\hat{\boldsymbol{\beta}}_{ols}$ in the next proposition.

Proposition 2 *The expected out-of-sample prediction error of $\hat{\boldsymbol{\beta}}_{ols}$ is*

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_{ols}) = \frac{\sigma^2}{n-p-1} (n-1 + \theta^2(1-a(1+\theta^2))) \approx \sigma^2 \frac{n-1}{n-p-1}. \quad (10)$$

Notice that the OLS prediction error diverges as p approaches $n-1$. To improve upon OLS, we consider the following family of linear shrinkage estimators of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(c, \delta) \equiv (1-\delta)\hat{\boldsymbol{\beta}}_{ols} + \delta c \mathbf{1}_p, \quad (11)$$

where $\delta \in [0, 1]$ is the *shrinkage intensity* and $c \in \mathbb{R}$ is the *shrinkage target*. This estimator shrinks OLS coefficients toward a constant target, and choosing $c = 0$ mimics the effect of the ridge penalty. We now use (9)–(11) to derive the prediction error of $\hat{\boldsymbol{\beta}}(c, \delta)$, the resulting oracle estimators of c and δ , and the prediction error reduction relative to OLS.

Proposition 3 *The following holds concerning the shrinkage estimator $\hat{\boldsymbol{\beta}}(c, \delta)$ in (11):*

1. *The expected out-of-sample prediction error of $\hat{\boldsymbol{\beta}}(c, \delta)$ is*

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{\beta}}(c, \delta)) &= \sigma^2 + \delta^2 ((\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\mu})^2 + \delta^2 (\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\Sigma}(\boldsymbol{\beta} - c \mathbf{1}_p) \\ &\quad + (1-\delta)^2 \frac{\sigma^2}{n-p-1} (p + \theta^2(1-a(1+\theta^2))). \end{aligned} \quad (12)$$

2. *For any δ , the oracle shrinkage target c^* minimizing $\mathcal{L}(\hat{\boldsymbol{\beta}}(c, \delta))$ is*

$$c^* = \frac{\mathbf{1}_p^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + (\mathbf{1}_p^\top \boldsymbol{\mu})(\boldsymbol{\beta}^\top \boldsymbol{\mu})}{\mathbf{1}_p^\top \boldsymbol{\Sigma} \mathbf{1}_p + (\mathbf{1}_p^\top \boldsymbol{\mu})^2}. \quad (13)$$

3. *Given c , the oracle shrinkage intensity δ^* minimizing $\mathcal{L}(\hat{\boldsymbol{\beta}}(c, \delta))$ is*

$$\begin{aligned} \delta^* &= \frac{\sigma^2 \frac{p+\theta^2(1-a(1+\theta^2))}{n-p-1}}{\sigma^2 \frac{p+\theta^2(1-a(1+\theta^2))}{n-p-1} + ((\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\mu})^2 + (\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\Sigma}(\boldsymbol{\beta} - c \mathbf{1}_p)} \\ &\approx \frac{\sigma^2 \frac{p}{n-p-1}}{\sigma^2 \frac{p}{n-p-1} + ((\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\mu})^2 + (\boldsymbol{\beta} - c \mathbf{1}_p)^\top \boldsymbol{\Sigma}(\boldsymbol{\beta} - c \mathbf{1}_p)}. \end{aligned} \quad (14)$$

4. *Given c , the optimal expected out-of-sample prediction error is*

$$\mathcal{L}(\hat{\boldsymbol{\beta}}(c, \delta^*)) = \mathcal{L}(\hat{\boldsymbol{\beta}}_{ols}) - \delta^* \sigma^2 \frac{p + \theta^2(1-a(1+\theta^2))}{n-p-1}, \quad (16)$$

which is smaller than $\mathcal{L}(\hat{\boldsymbol{\beta}}_{ols})$ in (10).

Several comments are in order. First, Proposition 3 shows that it is always optimal to shrink OLS coefficients to reduce

out-of-sample prediction error: δ^* is greater than zero and converges to zero as $n \rightarrow \infty$. Second, the optimal δ^* hits a tradeoff between estimation error, $\sigma^2 p/(n-p-1)$, and misspecification error that arises from bias, $((\beta - c\mathbf{1}_p)^\top \mu)^2$, and variance, $(\beta - c\mathbf{1}_p)^\top \Sigma (\beta - c\mathbf{1}_p)$. Third, c^* has natural properties: it is equal to k when the true $\beta = k\mathbf{1}_p$, and it is equal to the average of β in the orthonormal setting, i.e., $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$.

3. Consistent estimator of the shrinkage intensity

We now propose a feasible consistent estimator of the oracle shrinkage intensity δ^* in (14), which depends on four unobservable quantities: σ^2 , $(\beta - c\mathbf{1}_p)^\top \mu$, $(\beta - c\mathbf{1}_p)^\top \Sigma (\beta - c\mathbf{1}_p)$, and θ^2 . The estimates of these quantities we propose also allow us to estimate the optimal target c^* in (13) consistently. We derive these estimators under the same assumptions as in Section 2, i.e., the sample data are i.i.d. draws from $\mathbf{x} \sim N(\mu, \Sigma)$ and $\epsilon \sim N(0, \sigma^2)$, which are independent of each other.

First, we estimate the noise standard deviation as $\hat{\sigma}^2 \equiv \frac{1}{n} \epsilon^\top \epsilon \sim \frac{1}{n} \sigma^2 \chi_n^2$, which is unbiased.

Second, we estimate $(\beta - c\mathbf{1}_p)^\top \mu$ via its plug-in estimator $(\hat{\beta}_{ols} - c\mathbf{1}_p)^\top \hat{\mu}$, where $\hat{\mu} \equiv \frac{1}{n} \mathbf{X}^\top \mathbf{1}_n \sim N(\mu, \frac{1}{n} \Sigma)$ is the sample estimator of μ . This estimator of $(\beta - c\mathbf{1}_p)^\top \mu$ is unbiased because $\hat{\mu}$ is unbiased and because $\hat{\beta}_{ols}^\top \hat{\mu}$ is unbiased too since

$$\hat{\beta}_{ols}^\top \hat{\mu} = \beta^\top \hat{\mu} + \frac{1}{n} \epsilon^\top \mathbf{H} \mathbf{1}_n, \quad (17)$$

where $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is the hat matrix, and $\mathbb{E}[\epsilon^\top \mathbf{H} \mathbf{1}_n] = 0$.

Third, the quantity $(\beta - c\mathbf{1}_p)^\top \Sigma (\beta - c\mathbf{1}_p)$ requires estimating three terms:

$$(\beta - c\mathbf{1}_p)^\top \Sigma (\beta - c\mathbf{1}_p) = \beta^\top \Sigma \beta - 2c\beta^\top \Sigma \mathbf{1}_p + c^2 \mathbf{1}_p^\top \Sigma \mathbf{1}_p. \quad (18)$$

Let $\hat{\Sigma} \equiv \frac{1}{n-1} (\mathbf{X} - \mathbf{1}_n \hat{\mu}^\top)^\top (\mathbf{X} - \mathbf{1}_n \hat{\mu}^\top) \sim \frac{1}{n-1} W_p(n-1, \Sigma)$ be the sample covariance matrix, where W_p is a Wishart distribution of dimension p . We rely on the plug-in estimators of $\mathbf{1}_p^\top \Sigma \mathbf{1}_p$ and $\beta^\top \Sigma \mathbf{1}_p$, $\mathbf{1}_p^\top \hat{\Sigma} \mathbf{1}_p$ and $\hat{\beta}_{ols}^\top \hat{\Sigma} \mathbf{1}_p$, because one can easily show that they are unbiased. In contrast, we show in the next proposition that the plug-in estimator of $\beta^\top \Sigma \beta$, $\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}$, is biased.

Proposition 4 Consider the plug-in estimator of $\beta^\top \Sigma \beta$,

$$\begin{aligned} \hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols} &= \beta^\top \hat{\Sigma} \beta + \frac{2}{n-1} \epsilon^\top \mathbf{X} \beta - \frac{2}{n} (\beta^\top \hat{\mu}) (\epsilon^\top \mathbf{H} \mathbf{1}_n) \\ &\quad + \frac{1}{n-1} \epsilon^\top \mathbf{H} \epsilon - \frac{1}{n^2} (\epsilon^\top \mathbf{H} \mathbf{1}_n)^2, \end{aligned} \quad (19)$$

where $\mathbf{H}$ is the hat matrix. The bias of $\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}$ is

$$\begin{aligned} \text{Bias}[\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}] &\equiv \mathbb{E}[\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}] - \beta^\top \Sigma \beta \\ &= \frac{\sigma^2}{n} (p-1 + 1/(n-1) + a(1-p/n)), \end{aligned} \quad (20)$$

where a is defined in (7).

Because $\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}$ is biased, we estimate $\beta^\top \Sigma \beta$ as

$$\max(\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols} - \widehat{\text{Bias}}[\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}], 0),$$

where the estimated bias is obtained by plugging $\hat{\sigma}^2$ and the estimate of θ^2 below into (20).

Fourth, we estimate θ^2 in (3) via the adjusted estimator proposed by [4], which adjusts the unbiased estimate to avoid negative values. Provided $n > p+2$, it is defined as

$$\hat{\theta}_a^2 \equiv \frac{(n-p-2)\hat{\sigma}^2 - p}{n} + \frac{2(\hat{\theta}^2)^{p/2}(1+\hat{\theta}^2)^{-(n-2)/2}}{n \times B_{\hat{\theta}^2/(1+\hat{\theta}^2)}(p/2, (n-p)/2)}, \quad (21)$$

where $\hat{\theta}^2 \equiv \frac{n}{n-1} \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu}$ and $B_x(a, b) \equiv \int_0^x y^{a-1} (1-y)^{b-1} dy$ is the incomplete beta function.

4. Simulation results

We now test the prediction accuracy of our shrinkage estimator using simulated data with $p = 10, 25, 50, 100$, and 200. The sample size is either $n = 2p$ or $n = 5p$. We consider a noise standard deviation of $\sigma = 2$, and thus the minimum achievable expected out-of-sample prediction error is $\sigma^2 = 4$. Concerning the true moments of the random predictors $\mathbf{x}$, we consider two different settings. First, the orthonormal setting with $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$, in which case the optimal ridge penalty coefficient is known analytically. Second, a more general case in which predictors are correlated and non-standardized as in our theory.

4.1. Orthonormal predictors

We begin by comparing our shrinkage estimator $\hat{\beta}(c, \delta)$ in (11) with the classical ridge estimator:

$$\hat{\beta}_R \equiv (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}. \quad (22)$$

In general, the λ minimizing the expected out-of-sample prediction error is not known, which does not allow for a common ground comparison with our analytical shrinkage estimator. Therefore, we consider the special case in which predictors are orthonormal, i.e., $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$, because [8, Lemma 2] show that the optimal λ in this case is known analytically:

$$\lambda^* = p\sigma^2/(\beta^\top \beta). \quad (23)$$

The optimal λ^* is unfeasible because it depends on σ^2 and β . We estimate σ^2 by $\hat{\sigma}^2$ defined in Section 3. Concerning $\beta^\top \beta$, we derive the bias of the plug-in estimator $\hat{\beta}_{ols}^\top \hat{\beta}_{ols}$ in the next proposition, which is easily found because $\hat{\beta}_{ols}$ is unbiased and its covariance matrix in (6) is $\frac{\sigma^2}{n-p-1} \mathbf{I}_p$ when $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$.

Proposition 5 When $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$, the bias of $\hat{\beta}_{ols}^\top \hat{\beta}_{ols}$ is $\sigma^2 p/(n-p-1)$.

Therefore, we estimate λ^* in (23) by

$$\hat{\lambda}^* = \frac{p\hat{\sigma}^2}{\max(\hat{\beta}_{ols}^\top \hat{\beta}_{ols} - \frac{\hat{\sigma}^2 p}{n-p-1}, 0)}. \quad (24)$$

We compare this analytical ridge estimator with our shrinkage estimator $\hat{\beta}(c, \delta)$ in (11). We set $c = 0$, which is in line with the ridge estimator that implicitly shrinks OLS coefficients toward zero. When $\mu = \mathbf{0}_p$, $\Sigma = \mathbf{I}_p$, and $c = 0$, δ^* in (14) is

$$\delta^* = \frac{\frac{\sigma^2 p}{n-p-1}}{\frac{\sigma^2 p}{n-p-1} + \beta^\top \beta} = \frac{\lambda^*}{\lambda^* + n - p - 1}, \quad (25)$$

which we estimate using $\hat{\lambda}^*$ in (24). The higher the ridge penalty coefficient λ^* , the higher the shrinkage intensity δ^* .

Regarding the true regression coefficients β that we aim to estimate, we consider a sparse vector, i.e., only a limited number of predictors help predict the response. Specifically, 80% of the entries of β are zero and the 20% remaining entries are drawn uniformly between -1 and 1 .

Next, we proceed by Monte Carlo simulations to compare the expected out-of-sample prediction error of several estimators of β . Specifically, given p , we draw $M = 10,000$ times n data from model (4). Then, for each draw $m = 1, \dots, M$, we obtain the estimated regression coefficients $\hat{\beta}_m$ and approximate the expected out-of-sample prediction error as

$$\mathcal{L}(\hat{\beta}) \approx \frac{1}{M} \sum_{m=1}^M \sigma^2 + ((\beta - \hat{\beta}_m)^\top \mu)^2 + (\beta - \hat{\beta}_m)^\top \Sigma (\beta - \hat{\beta}_m) \quad (26)$$

$$= \sigma^2 + \frac{1}{M} \sum_{m=1}^M (\beta - \hat{\beta}_m)^\top (\beta - \hat{\beta}_m) \quad (27)$$

because $\mu = \mathbf{0}_p$ and $\Sigma = \mathbf{I}_p$. We can evaluate $\mathcal{L}(\hat{\beta})$ because we know the true value of the parameters in our simulation setting. Note that σ^2 , β , μ , and Σ remain constant for each simulation.

We consider the following estimators $\hat{\beta}$. First, the unfeasible true β , which yields a prediction error of $\sigma^2 = 4$. Second, the OLS estimator $\hat{\beta}_{ols}$, which, from Proposition 2, has a prediction error of $\sigma^2 \frac{n-1}{n-p-1}$. Third and fourth, our proposed shrinkage estimator (11) with $c = 0$ and using either the true (but unfeasible) shrinkage intensity δ^* in (25) or its estimated feasible version, which allows us to assess the impact of estimation errors in the shrinkage intensity. The true intensity δ^* yields the prediction error in (16) with $\theta^2 = 0$ since $\mu = \mathbf{0}_p$. Fifth and sixth, the ridge estimator $\hat{\beta}_R$ in (22) using the true (but unfeasible) penalty coefficient λ^* in (23) or its estimated feasible version in (24).

The results are in Table 1. Ignoring estimation errors in the optimal δ^* and λ^* , we find that shrinking OLS coefficients to zero as we propose yields a larger prediction error than ridge, but only slightly so. The only non-negligible losses are when p is large and n is small. Then, the loss incurred from estimation errors in δ^* and λ^* depends essentially on n rather than p , because these are univariate quantities. Therefore, this loss is small when $n = 5p$ for all p considered, but can get more substantial when p is not large and $n = 2p$. Finally, our method substantially improves upon OLS. For example, when $p = 25$ and $n = 2p$, OLS yields a prediction error of 8.17, while our method yields 5.29 optimally and 5.56 when δ^* is estimated, which is much closer to the minimum error of 4.

Overall, therefore, our shrinkage methodology significantly improves upon OLS and delivers nearly the same performance

Table 1: Comparison of out-of-sample prediction error with analytical ridge when predictors are orthonormal.

This table reports the expected out-of-sample prediction error of several estimators of linear regression coefficients following the methodology described in Section 4.1. We consider a setting with orthonormal predictors ($\mu = \mathbf{0}_p$, $\Sigma = \mathbf{I}_p$), in which case the optimal ridge penalty coefficient is known analytically in (23). p is the number of predictors and n is the sample size. Min error is the variance of the noise term. OLS is ordinary least squares. Analytical ridge is either with the optimal penalty coefficient λ^* in (23) or the feasible estimated one in (24). Shrink OLS to $c = 0$ is our analytical shrinkage estimator in (11) that shrinks OLS coefficients to a constant zero vector, either with the optimal shrinkage intensity δ^* in (14) or the feasible estimated one in Section 3.

		Min error	OLS	Analytical ridge		Shrink OLS to $c = 0$	
				Optimal	Feasible	Optimal	Feasible
$p = 10$	$n = 2p$	4	8.44	4.20	4.41	4.21	4.66
	$n = 5p$	4	5.03	4.18	4.24	4.18	4.26
$p = 25$	$n = 2p$	4	8.17	5.08	5.28	5.29	5.56
	$n = 5p$	4	5.01	4.62	4.65	4.66	4.70
$p = 50$	$n = 2p$	4	8.08	5.45	5.53	5.76	5.89
	$n = 5p$	4	5.00	4.72	4.74	4.76	4.77
$p = 100$	$n = 2p$	4	8.04	6.02	6.04	6.43	6.48
	$n = 5p$	4	5.00	4.84	4.84	4.86	4.87
$p = 200$	$n = 2p$	4	8.02	6.76	6.77	7.17	7.18
	$n = 5p$	4	5.00	4.93	4.93	4.94	4.94

as that of the ridge penalty, but with the added benefit that, unlike ridge, our shrinkage estimator is known analytically for any moments of predictors μ and Σ .

4.2. Correlated and non-standardized predictors

We now perform a similar analysis to that in Section 4.1 but in the more realistic setting in which predictors are not orthonormal, which our theory can also accommodate. Specifically, for each considered value of p , we draw each entry of μ uniformly between -1 and 1 . For the covariance matrix Σ , each standard deviation is drawn uniformly between 0.1 and 1.9 and thus is equal to half the noise standard deviation σ on average. We then set the correlation matrix randomly following [11].

We apply our shrinkage method that shrinks OLS coefficients toward a constant target. We consider two choices of shrinkage target c . First, $c = 0$, which is a natural choice absent any particular information. Second, the optimal value c^* in (13) that minimizes the expected out-of-sample prediction error. For both choices of c , we consider the feasible estimator and the optimal one that assumes the true δ^* and c^* are known. This allows us to assess whether the optimal c^* , which delivers smaller out-of-sample prediction error than $c = 0$ when it is known, also outperforms when it is estimated.

Concerning the true β , we study two cases. First, the sparse specification in Section 4.1, which may favor setting $c = 0$ because 80% of the entries of β are zero and the remaining ones are zero on average. Therefore, we consider a second specification in which $c = 0$ is wrong because the entries of β are drawn uniformly between 0 and 2, and thus, are one on average.

We compare our linear shrinkage estimator with that of [9], which is a Stein-rule estimator that linearly combines OLS and PC regression (PCR) to minimize the expected out-of-sample prediction error. Specifically, let V_k be the $p \times k$ matrix that

contains the first k eigenvectors of $X^\top X$. Define the OLS regression coefficients on the first k PCs as $\gamma_k = (W_k^\top W_k)^{-1} W_k^\top y$ where $W_k = X V_k$. Then, the PCR estimator of β is $\hat{\beta}_k = V_k \gamma_k$. [9] derive a Stein-rule combination of $\hat{\beta}_k$ and $\hat{\beta}_{ols}$ that achieves a lower prediction error than OLS for any $k < p - 2$:

$$\hat{\beta}_S = (1 - \delta_S) \hat{\beta}_{ols} + \delta_S \hat{\beta}_k, \quad (28)$$

where

$$\delta_S = \frac{(p - k - 2) \tilde{\sigma}^2}{\|X \hat{\beta}_{ols} - X \hat{\beta}_k\|^2} \quad (29)$$

and $\tilde{\sigma}^2 = (n - p + 1)^{-1} \|y - X \hat{\beta}_{ols}\|^2$. We cap δ_S between zero and one and select k so that 80% of the variance is retained.

Finally, we approximate the expected out-of-sample prediction error of the different estimators of β by (26).

Table 2: Out-of-sample prediction error for correlated and non-standardized predictors.

This table reports the expected out-of-sample prediction error of several estimators of linear regression coefficients following the methodology described in Section 4.2. We consider a setting with correlated and non-standardized predictors. p is the number of predictors and n is the sample size. Min error is the variance of the noise term. OLS is ordinary least squares. Oman is the shrinkage estimator of [9] in (28)–(29). Shrink OLS to $c = 0$ and $c = c^*$ is our analytical shrinkage estimator in (11) that shrinks OLS coefficients either to $c = 0$ or to $c = c^*$ in (13), using the optimal shrinkage intensity δ^* in (14) and target c^* or the feasible estimated ones in Section 3. Panel A considers the sparse β specification from Section 4.1 and panel B the specification from Section 4.2 in which each $\beta_i = 1$ on average.

		Min error	OLS	Oman	Shrink OLS to $c = 0$ Optimal	Shrink OLS to $c = 0$ Feasible	Shrink OLS to c^* Optimal	Shrink OLS to c^* Feasible
A) Sparse β								
$p = 10$	$n = 2p$	4	8.44	5.77	4.11	4.18	4.11	4.36
	$n = 5p$	4	5.03	4.48	4.10	4.15	4.10	4.22
$p = 25$	$n = 2p$	4	8.17	5.63	5.82	5.87	5.44	5.59
	$n = 5p$	4	5.01	4.67	4.77	4.79	4.69	4.74
$p = 50$	$n = 2p$	4	8.08	6.04	6.44	6.47	6.22	6.30
	$n = 5p$	4	5.01	4.78	4.86	4.87	4.83	4.85
$p = 100$	$n = 2p$	4	8.04	6.81	7.14	7.15	7.14	7.17
	$n = 5p$	4	5.00	4.89	4.94	4.94	4.94	4.94
$p = 200$	$n = 2p$	4	8.02	6.96	7.53	7.53	7.48	7.49
	$n = 5p$	4	5.00	4.91	4.97	4.97	4.96	4.97
B) $\beta_i = 1$ on average								
$p = 10$	$n = 2p$	4	8.44	7.22	7.43	7.54	4.74	5.06
	$n = 5p$	4	5.03	4.93	4.96	4.98	4.48	4.63
$p = 25$	$n = 2p$	4	8.17	7.78	7.99	8.00	7.08	7.19
	$n = 5p$	4	5.01	4.98	5.00	5.00	4.93	4.95
$p = 50$	$n = 2p$	4	8.08	7.81	7.95	7.96	7.54	7.57
	$n = 5p$	4	5.01	4.98	5.00	4.99	4.97	4.97
$p = 100$	$n = 2p$	4	8.04	7.84	7.96	7.96	7.78	7.78
	$n = 5p$	4	5.00	4.99	5.00	5.00	4.99	4.99
$p = 200$	$n = 2p$	4	8.02	7.96	7.99	8.00	7.87	7.89
	$n = 5p$	4	5.00	5.00	5.00	5.00	4.99	4.99

The results are in Table 2. We make four main observations. First, our proposed methodology continues to substantially improve upon OLS.

Second, for both specifications of β , the optimal shrinkage target c^* delivers a smaller prediction error than $c = 0$, even if c^* is estimated. The gain is particularly large when the entries of β are one on average, and thus setting $c = 0$ is wrong.

Third, the prediction errors obtained with our estimator in panel A of Table 2 are comparable to those in the orthonormal

setting in Table 1, which means that our approach also performs well in a more complicated setting where predictors are correlated and heterogeneous. To understand this, consider the case in which c and δ^* are known without error. Then, the expected out-of-sample prediction error with our approach in (16) is, under the very accurate approximation $a \approx 1/(1 + \theta^2)$,

$$\mathcal{L}(\hat{\beta}(c, \delta^*)) = \sigma^2 \frac{n - \delta^* p - 1}{n - p - 1}. \quad (30)$$

Only δ^* changes across datasets with different μ and Σ , and the δ^* we obtain in the setting of Table 1 and of panel A in Table 2 are not substantially different.

Fourth, our approach offers a competitive prediction performance relative to the Oman estimator of [9]. In panel B where $\beta_i = 1$ on average, our proposed estimator consistently outperforms. In panel A where β is sparse, the two estimators perform similarly when $n = 5p$ because they both find it optimal not to shrink OLS intensively. When $n = 2p$, our proposed estimator outperforms when p is rather small ($p = 10$ and 25), and the Oman estimator outperforms otherwise. Finally, Table 2 shows that in general the difference between the two approaches is not large, which points to a general robustness of linear shrinkage.

5. Conclusion

We introduce a shrinkage methodology for linear regression designed to minimize the expected out-of-sample prediction error and solvable analytically for any moments of predictors. The regression method we propose directly and optimally shrinks OLS coefficients to a constant target. This formulation is an alternative to the standard approach that penalizes different norms of regression coefficients, for which the optimal penalty coefficient is not available in closed form in general. Simulation results show that, in the setting with orthonormal predictors in which the optimal ridge penalty coefficient is known in closed form, our approach often delivers nearly the same performance. Moreover, in the more general setting with correlated and non-standardized predictors, which our analytical approach can also accommodate, we obtain a competitive performance relative to the linear shrinkage estimator of [9].

6. Proofs of results

6.1. Proof of Proposition 1

Let $\hat{\mu} \equiv \frac{1}{n} X^\top \mathbf{1}_n \sim N(\mu, \frac{1}{n} \Sigma)$ be the sample mean of x and $\hat{\Sigma} \equiv \frac{1}{n} (X - \mathbf{1}_n \hat{\mu}^\top)^\top (X - \mathbf{1}_n \hat{\mu}^\top) \sim \frac{1}{n} W_p(n - 1, \Sigma)$ the sample covariance matrix, where W_p is a Wishart distribution of dimension p . Note that for $\hat{\Sigma}$ we use a $1/n$ coefficient instead of the $1/(n - 1)$ we use in the main text. Then, from the first equality in (6),

$$\text{Var}[\hat{\beta}_{ols}] = \frac{\sigma^2}{n} \mathbb{E}[(\hat{\Sigma} + \hat{\mu} \hat{\mu}^\top)^{-1}]. \quad (31)$$

We now derive a formula for $\mathbb{E}[(\hat{\Sigma} + \hat{\mu} \hat{\mu}^\top)^{-1}]$. Let $P = [v, P_1]$ be a $p \times p$ orthonormal matrix with $v = \Sigma^{-\frac{1}{2}} \mu / \theta$. Define

$$z = \sqrt{n} P^\top \Sigma^{-\frac{1}{2}} \hat{\mu} \sim N([\sqrt{n} \theta, \mathbf{0}_{p-1}^\top]^\top, I_p), \quad (32)$$

$$\mathbf{W} = n\mathbf{P}^\top \Sigma^{-\frac{1}{2}} \hat{\Sigma} \Sigma^{-\frac{1}{2}} \mathbf{P} \sim W_p(n-1, \mathbf{I}_p), \quad (33)$$

which are independent of each other. Using $\mathbf{z}$ and $\mathbf{W}$, we have

$$\begin{aligned} \hat{\Sigma} + \hat{\mu}\hat{\mu}^\top &= \frac{1}{n}\Sigma^{\frac{1}{2}}\mathbf{P}\mathbf{W}\mathbf{P}^\top\Sigma^{\frac{1}{2}} + \frac{1}{n}\Sigma^{\frac{1}{2}}\mathbf{P}\mathbf{z}\mathbf{z}^\top\mathbf{P}^\top\Sigma^{\frac{1}{2}} \\ &= \frac{1}{n}\Sigma^{\frac{1}{2}}\mathbf{P}(\mathbf{W} + \mathbf{z}\mathbf{z}^\top)\mathbf{P}^\top\Sigma^{\frac{1}{2}}. \end{aligned} \quad (34)$$

It follows that

$$(\hat{\Sigma} + \hat{\mu}\hat{\mu}^\top)^{-1} = n\Sigma^{-\frac{1}{2}}\mathbf{P}(\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1}\mathbf{P}^\top\Sigma^{-\frac{1}{2}}, \quad (35)$$

so we can focus on obtaining $\mathbb{E}[(\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1}]$. Note that

$$\tilde{\mathbf{W}} \equiv \mathbf{W} + \mathbf{z}\mathbf{z}^\top \sim W_p(n, \mathbf{I}_p, \mathbf{\Lambda}) \quad (36)$$

has a noncentral Wishart distribution with n degrees of freedom, a covariance matrix $\mathbf{I}_p$, and a noncentrality matrix $\mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $[\theta^2, \mathbf{0}_{p-1}^\top]$.

Using [2, Theorem 1], $\mathbb{E}[\tilde{\mathbf{W}}^{-1}]$ is a diagonal matrix whose diagonal elements are equal to $1/(n-p-1)$ except for its (1,1) element that is equal to

$$\mathbb{E}[\tilde{\mathbf{W}}_{11}^{-1}] = \frac{1}{n-p-1} \sum_{k=0}^{\infty} \frac{\left(-\frac{n\theta^2}{2}\right)^k}{\left(\frac{n}{2}\right)_k} = \frac{1}{n-p-1} {}_1F_1\left(1; \frac{n}{2}; -\frac{n\theta^2}{2}\right), \quad (37)$$

where $(x)_k = x(x+1)\cdots(x+k-1)$ is the Pochhammer symbol. Putting all the terms together, we obtain

$$\mathbb{E}[(\hat{\Sigma} + \hat{\mu}\hat{\mu}^\top)^{-1}] = \frac{n}{n-p-1}(\Sigma^{-1} - a\Sigma^{-1}\mu\mu^\top\Sigma^{-1}), \quad (38)$$

where a is defined in (7), and plugging this identity into (31) yields the desired result. Finally, as $n \rightarrow \infty$, we have that

$$\mathbb{E}[(\hat{\Sigma} + \hat{\mu}\hat{\mu}^\top)^{-1}] \rightarrow (\Sigma + \mu\mu^\top)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}\mu\mu^\top\Sigma^{-1}}{1 + \theta^2}, \quad (39)$$

which means that $a \rightarrow 1/(1 + \theta^2)$ as $n \rightarrow \infty$.

6.2. Proof of Proposition 4

Using the formula for $\hat{\beta}_{ols}$ in (5) and $\hat{\Sigma} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X} - \hat{\mu}\hat{\mu}^\top$, $\hat{\beta}_{ols}^\top \hat{\Sigma} \hat{\beta}_{ols}$ simplifies to (19), which is composed of five terms. The expectation of the first term is $\beta^\top \Sigma \beta$ because $\hat{\Sigma}$ is unbiased. The expectation of the second and third terms is zero because ϵ has zero mean and is independent of $\mathbf{X}$. The expectation of the fourth term is $\sigma^2 p/(n-1)$ because $\text{Trace}(\mathbf{H}) = p$. Finally, the expectation of the fifth term is equal to the variance of $\frac{1}{n}\epsilon^\top \mathbf{H}\mathbf{1}_n$,

$$\begin{aligned} \mathbb{V}\text{ar}[n^{-1}\epsilon^\top \mathbf{H}\mathbf{1}_n] &= n^{-2}\mathbb{E}[\mathbb{V}\text{ar}[\epsilon^\top \mathbf{H}\mathbf{1}_n|\mathbf{H}]] \\ &= \sigma^2 n^{-2}\mathbb{E}[\mathbf{1}_n^\top \mathbf{H}\mathbf{1}_n] \\ &= \sigma^2 \mathbb{E}[\hat{\mu}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \hat{\mu}]. \end{aligned} \quad (40)$$

Now, using the random variables $\mathbf{z}$ and $\mathbf{W}$ with distributions given by (32) and (33), respectively, we have that

$$\mathbb{E}[\hat{\mu}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \hat{\mu}] = \frac{1}{n}\mathbb{E}[\mathbf{z}^\top (\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1} \mathbf{z}], \quad (41)$$

where

$$\mathbf{z}^\top (\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1} \mathbf{z} = \mathbf{z}^\top \left(\mathbf{W}^{-1} - \frac{\mathbf{W}^{-1} \mathbf{z} \mathbf{z}^\top \mathbf{W}^{-1}}{1 + \mathbf{z}^\top \mathbf{W}^{-1} \mathbf{z}} \right) \mathbf{z} = \frac{\mathbf{z}^\top \mathbf{W}^{-1} \mathbf{z}}{1 + \mathbf{z}^\top \mathbf{W}^{-1} \mathbf{z}}. \quad (42)$$

Since

$$\mathbf{z}^\top \mathbf{W}^{-1} \mathbf{z} \stackrel{d}{=} \frac{u}{v}, \quad (43)$$

where $u \sim \chi_p^2(n\theta^2) \perp v \sim \chi_{n-p}^2$, we have

$$\mathbf{z}^\top (\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1} \mathbf{z} = \frac{u}{u+v}, \quad (44)$$

which has a noncentral beta distribution with degrees of freedom $p/2$, $(n-p)/2$, and a noncentrality parameter $n\theta^2$. It follows that

$$\mathbb{E}[\mathbf{z}^\top (\mathbf{W} + \mathbf{z}\mathbf{z}^\top)^{-1} \mathbf{z}] = 1 - \frac{n-p}{n} {}_1F_1\left(1; \frac{n+2}{2}; -\frac{n\theta^2}{2}\right) = 1 - a \frac{n-p}{n}. \quad (45)$$

This means that the expectation of the fifth term in (19) is

$$\mathbb{E}[n^{-2}(\epsilon^\top \mathbf{H}\mathbf{1}_n)^2] = \frac{\sigma^2}{n}(1 - a(1 - p/n)). \quad (46)$$

Putting all the results together yields the bias formula in (20).

Declarations of interest None.

Acknowledgments I thank Raymond Kan (University of Toronto) for his help on proving the result in Proposition 1. This work was supported by the Fonds de la Recherche Scientifique (F.R.S.-FNRS) [grant number J.0115.22].

References

- [1] Ahmed, S.E., 2014. Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation. Springer.
- [2] Hillier, G., Kan, R., 2021. Properties of the inverse of a noncentral wishart matrix. *Econometric Theory* (forthcoming).
- [3] Hoerl, A., Kennard, R., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- [4] Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis* 42, 621–656.
- [5] Lassance, N., Martín-Utrera, A., Simaan, M., 2022. The risk of out-of-sample portfolio performance. Available at SSRN 3855546.
- [6] Lederer, J., 2021. *Fundamentals of high-dimensional statistics*. Springer.
- [7] Lederer, J., Müller, C., 2015. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the trex. *Proceedings of the 29th AAAI Conference on Artificial Intelligence* 29, 2729–2735.
- [8] Nakkiran, P., Venkat, P., Kakade, S., Ma, T., 2021. Optimal regularization can mitigate double descent. *International Conference on Learning Representations*.
- [9] Oman, S., 1991. Random calibration with many measurements: An application of Stein estimation. *Technometrics* 33, 187–195.
- [10] Phukongtong, S., Lisawadi, S., Ahmed, S.E., 2022. Penalty, post pretest and shrinkage strategies in a partially linear model. *Communications in Statistics - Simulation and Computation* 51, 6004–6025.
- [11] Pourahmadi, M., Wang, X., 2015. Distribution of random correlation matrices: Hyperspherical parameterization of the cholesky factor. *Statistics & Probability Letters* 106, 5–12.
- [12] Tay, J., Friedman, J., Tibshirani, R., 2021. Principal component-guided sparse regression. *The Canadian Journal of Statistics* 49, 1222–1257.
- [13] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* 58, 267–288.
- [14] Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* 67, 301–320.