A Simple Approximation to the Distribution of the Ridge Regression Estimator*

José Luis Montiel Olea Ryan Strong Amilcar Velez Zhuoheng Xu
Haomin Yu†

July 1, 2026

Abstract

We present a simple Gaussian approximation to the finite-sample distribution of the classical ridge regression estimator. Our approximation captures the fact that, in finite samples, the ridge regression estimator trades off bias and variance to improve estimation and prediction error. Our approximation is based on nonstandard asymptotics where *i*) we let the estimator's regularization parameter grow proportionally to the sample size; and *ii*) we treat the population regression coefficients as *local* to the reference vector that defines the estimator's direction of shrinkage. In contrast to other asymptotic approximations available in the literature, we allow for general forms of heteroskedasticity and autocorrelation in the data generating process (at the cost of considering a low-dimensional model where covariates are not allowed to grow with the sample size). We use our simple Gaussian approximation to propose two new strategies to select the regularization parameter for the ridge regression estimator. The suggested strategies select the regularization parameter to minimize either average or worst-case excess prediction risk, where risk is computed using our suggested Gaussian approximation.

1 Introduction

We have access to a dataset $D_n \equiv \{(y_i, x_i^\top)\}_{i=1}^n$ comprised of n observations of a real-valued outcome variable, $y_i \in \mathbb{R}$, and a vector of k covariates, $x_i \in \mathbb{R}^k$. The dataset D_n is assumed to have been

*We would like to thank Lihua Lei and Dacheng Xiu for very helpful comments and suggestions. Montiel Olea gratefully acknowledges financial support by the National Science Foundation Grant SES-2315600.

†All of the authors are affiliated to the Department of Economics at Cornell University. Corresponding author: montiel.olea@gmail.com

generated by the statistical model

$$y_i = x_i^\top \beta + \epsilon_i, \quad \{(x_i^\top, \epsilon_i)\}_{i=1}^n \sim \mathbb{P}. \tag{1}$$

We are willing to restrict the data generating processes under consideration by requiring that

$$(1/n)\sum_{i=1}^n x_i x_i^\top \xrightarrow{p} \Sigma, \quad \text{and} \quad (1/\sqrt{n})\sum_{i=1}^n x_i \epsilon_i \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Omega), \tag{2}$$

where both Σ and Ω are positive definite matrices. These assumptions allow for general forms of heteroskedasticity and autocorrelation in the data generating process.

We are interested in approximating the finite-sample distribution of the *ridge regression estimator:*

$$\widehat{\beta}_{\lambda_n} \equiv \left(\sum_{i=1}^n x_i x_i^\top + \lambda_n \mathbb{I}_k\right)^{-1} \left(\sum_{i=1}^n x_i y_i + \lambda_n \beta_0\right). \tag{3}$$

We refer to the nonnegative scalar λ_n as the *regularization parameter* and to $\beta_0 \in \mathbb{R}^k$ as the *reference vector*. We index the regularization parameter by the sample size to allow for the possibility that it depends on the dataset D_n.

Our first result (Theorem 1) shows that there is a sense in which, under (1)-(2), we can approximate the distribution $\widehat{\beta}_{\lambda_n} - \beta$ by the Gaussian distribution:

$$\mathcal{N}_k\left(-(\lambda_n/n)(\Sigma + (\lambda_n/n)\mathbb{I}_k)^{-1}(\beta - \beta_0), (1/n)(\Sigma + (\lambda_n/n)\mathbb{I}_k)^{-1}\Omega(\Sigma + (\lambda_n/n)\mathbb{I}_k)^{-1}\right). \tag{4}$$

Our simple approximation captures the fact that, in finite samples, the choice of regularization parameter and reference vector affect the bias and variance of the ridge regression estimator. We view the mean and variance in the limiting distribution in (4) as providing simple generalizations of common formulae for the bias and variance of the ridge estimator derived conditional on covariates and under i.i.d. sampling; see, for example, Equations 29.8 and 29.9 in Hansen (2022). The Gaussian

approximation in (4) can also be viewed as a generalization of the asymptotic distribution of the ridge estimator presented in the seminal work of Knight and Fu (2000), which they derive imposing conditional homoskedasticity (in our framework, heteroskedasticity and serial autocorrelation are captured in the matrix Ω).

In order to derive Theorem 1, we use *nonstandard asymptotics*.[1] As we will explain later, we assume that the true coefficient β in the model (1) (henceforth denoted β_n) depends on the sample size. We treat β_n as *local to the reference vector* β_0 that defines the estimator's direction of shrinkage. More concretely, we assume that:

$$\sqrt{n}(\beta_n - \beta_0) \to b, \tag{5}$$

for some $b \in \mathbb{R}^k$. Our asymptotic analysis focuses on (possibly data-dependent) choices of the regularization parameter that grow proportionally to the sample size, in the sense that

$$\lambda_n/n \xrightarrow{p} \lambda \in [0, \infty). \tag{6}$$

Our second result (Theorem 2) uses the Gaussian distribution in (4) to provide an approximation to the *excess prediction risk* of the ridge regression estimator. As we explain later, excess prediction risk is defined as the additional prediction risk relative to an oracle that knows β_n. Predictions of the outcome variable based on the ridge regression estimator take the form

$$\widehat{a}_{\lambda_n}(x) = x^\top \widehat{\beta}_{\lambda_n}, \tag{7}$$

where $\widehat{\beta}_{\lambda_n}$ is defined in (3).

The exact finite-sample analysis of the prediction risk of ridge regression is challenging even under a stylized homoskedastic, Gaussian regression model. There are different recent papers in the statistics literature that provide approximations (and lower/upper bounds) to this prediction

[1] See Powell (2017) for a description of the role that nonstandard asymptotics plays in modern econometrics.

risk. For example, Dobriban and Wager (2018), Hastie, Montanari, Rosset, and Tibshirani (2022), Mourtada and Rosasco (2022), Atanasov, Zavatone-Veth, and Pehlevan (2024). We show that it is possible to use our Theorem 1 to provide an approximation to the excess prediction risk of ridge regression as a function of three types of parameters: λ (the probability limit of λ_n/n); b (the parameter controlling the local-to-β_0 regression coefficient β_n); and the variance parameters Σ, Ω. Importantly, the variance parameters can be consistently estimated, but the local parameter b cannot.

Our third result (Theorem 3) presents closed-form, data-driven recommendations for the selection of the regularization parameter λ_n. Our formulae are designed for the special case in which—under the true unknown data generating process—the covariates are asymptotically uncorrelated and have the same variance; that is, $\Sigma = \sigma_x^2 \mathbb{I}_k$, where $\mathbb{I}_k$ is the identity matrix of dimension k and σ_x^2 is a common variance parameter. Following Hastie et al. (2022), we refer to this scenario as one having *isotropic features*.[2] The key step in deriving our closed-form formulae is to choose the regularization parameter to optimize the excess prediction risk approximation in Theorem 2. As we discussed before, this approximation depends on the local parameter b and variance parameters σ_x^2 and Ω. While the latter parameters can typically be consistently estimated under mild assumptions, the local parameter b cannot. We use the common decision-theoretic principles of average and worst-case risk (see Ferguson (1967)) to account for the unknown parameter b in the approximate excess prediction risk. If π is a prior distribution on b such that $\mathbb{E}_\pi[b^\top b] < \infty$ and $\mathbb{E}_\pi[b^\top b] > 0$, Theorem 3 shows that the π-optimal selection of the regularization parameter λ_n in ridge regression is

$$\widehat{\lambda}^*_{n,\pi} = n \cdot \frac{\text{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 \mathbb{E}_\pi[b^\top b]}. \tag{8}$$

When the localization parameter b admits a known bound B on its norm ($\|b\| \leq B$), the selection of the regularization parameter λ_n that minimizes approximate worst-case excess risk (and that we

[2]Although we note that in Hastie et al. (2022) Σ equals the identity matrix, whereas we only require $\Sigma = \sigma_x^2 \mathbb{I}_k$.

term the *minimax* choice of λ_n) is:

$$\widehat{\lambda}^*_{n,\text{minimax}} = n \cdot \frac{\text{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 B^2}. \tag{9}$$

While the formulae in (8)-(9) pertain to the case of isotropic features, the idea of optimizing approximate excess prediction risk (either by using an average or worst-case criterion) is more general. We explain how to operationalize both of these approaches for nonisotropic features; see Sections 3.3.1 and 3.3.2. In both cases, the optimal choice of regularization parameter is the solution of a nonconvex optimization problem over the positive part of the real line. In this case, we suggest to choose λ_n by evaluating the objective functions on a grid of candidate regularization parameters. This approach is the same as the one used by conventional statistical packages to find and implement the usual cross-validated choice of regularization parameter.

Related Literature: The introduction of the ridge estimator to regression analysis is often credited to Hoerl and Kennard (1970), although the estimator is based on the earlier work of Hoerl (1962); see Hoerl (2020). The textbook version of the ridge regression estimator uses a reference vector $\beta_0 = 0_{k\times 1}$; see, for example, Chapter 3.4, Equation 3.41 in Friedman, Hastie, and Tibshirani (2017) or Chapter 29.5 in Hansen (2022). The more general formulation given in equation (3) has several precedents in the literature; for example, Equation 2.1 in Swindel (1976). See also Anup and Maddala (1984).

The high-level assumptions in (2) that we use to analyze the distribution of the ridge estimator are common in the econometrics literature. For example, it is well known that (2) can be verified under i.i.d. sampling and standard primitive conditions on the joint distribution of x_i and ϵ_i (see Assumption 7.2 in Hansen (2022) and Theorems 7.1 and 7.2 therein). However, our framework allows us to incorporate richer data structures, where it is possible to depart from i.i.d. sampling and have general forms of heteroskedasticity and autocorrelation in $(x_i^\top, \epsilon_i)$.

The approximation we propose in (4) has not—to the best of our knowledge—appeared before in

the literature expressed at this level of generality. The closest reference that we are aware of is the seminal work of Knight and Fu (2000), who present a Gaussian approximation to the distribution of the ridge regression estimator assuming independence between x_i and ϵ_i (and independence across observations). The distribution in (4) coincides with their result when specialized to the case in which Ω is conditionally homoskedastic (that is, $\Omega = \mathbb{E}[\epsilon_i^2]\Sigma$) and $\beta_0 = 0_{k\times 1}$. We view the mean and variance in the limiting distribution in (4) as providing simple generalizations of common formulae for the bias and variance of the ridge estimator derived conditional on covariates and under i.i.d. sampling; see, for example, Equations 29.8 and 29.9 in Hansen (2022) and also Section 4 in Hoerl and Kennard (1970).

The idea of considering drifting sequences of parameter values—as those in (5) and (6)—in order to improve the distributional approximations provided by standard asymptotic theory has a long history in econometrics. See Powell (2017) for examples and details on the use of nonstandard asymptotics in econometrics. More concretely, the idea of using nonstandard asymptotics to analyze the distribution of the ridge regression estimator also appears in Knight and Fu (2000) (they refer to it as *triangular array asymptotics*), and more recently in the work of Shen and Xiu (2025), who assess the predictive performance of several machine learning methods in high-dimensional regressions with low signal-to-noise ratios.

There is also a recent literature analyzing the risk of predictions based on the ridge regression estimator, where the outcome variable predictions take the same form as in (7). For example, Hsu, Kakade, and Zhang (2012), Dobriban and Wager (2018), Hastie et al. (2022), Mourtada and Rosasco (2022), Atanasov et al. (2024). All these papers make distributional assumptions that preclude data generating processes that exhibit at least one of the following features: conditional heteroskedasticity, time-series autocorrelation, or a nonlinear conditional expectation function. It is important to mention, however, that the additional assumptions in these papers allow them to consider *high-dimensional* approximations where the number of covariates can be large relative to the sample size. In contrast, all the analysis in our paper pertains to a *low-dimensional* model

where we treat the number of covariates as fixed, and let the sample size diverge to infinity.

Finally, the idea of using an approximation of the risk function to choose regularization parameters has a long precedent in statistics and econometrics; see, for example the discussion of Abadie and Kasy (2019) on p. 744. To the best of our knowledge the formulae we provide in (8) and (9) are new to the literature. It is important to mention, however, that one of the initial motivations of our paper was to recover an expression analogous to the optimal regularization parameter for ridge regression given by Hastie et al. (2022) in their Corollary 6 to Theorem 6. Their results show that using high-dimensional asymptotics where $k/n \to \gamma$, the oracle choice of regularization parameter for the ridge estimator (assuming x_i and ϵ_i are independent and that $\Sigma = \mathbb{I}_k$) takes the form $\lambda^* = n\mathbb{E}[\epsilon_i^2]\gamma/\|\beta\|^2$. Assuming conditional homoskedasticity, our formulae in (8) and (9) become analogous to Hastie et al. (2022)'s formula, but the true unknown β_n is replaced by either the prior mean of $\|b\|^2$ or the maximum value of $\|b\|^2$.

Outline: The rest of this paper is organized as follows. Section 2 introduces notation, framework, and main assumptions. Section 3 presents our main results. Section 4 presents simulation evidence to illustrate and support our main results. Section 5 concludes.

2 Notation and Framework

An econometrician has access to a dataset $D_n \equiv \{(y_i, x_i^\top)\}_{i=1}^n$ comprised of n observations of a real-valued outcome variable, $y_i \in \mathbb{R}$, and a vector of k covariates, $x_i \in \mathbb{R}^k$. The dataset D_n is assumed to have been generated by the statistical model

$$y_i = x_i^\top \beta_n + \epsilon_i, \quad \{(x_i^\top, \epsilon_i)\}_{i=1}^n \sim \mathbb{P}_n. \tag{10}$$

The parameters of this statistical model are i) the unknown vector of slope coefficients, $\beta_n \in \mathbb{R}^k$, and ii) the unknown distribution of the tuple $((x_1^\top, \epsilon_1), \ldots, (x_n^\top, \epsilon_n))$, which we denote as $\mathbb{P}_n$. Note

that we have indexed both β_n and $\mathbb{P}_n$ by the sample size n. While this does not make a difference when conducting finite-sample analysis (since, in that case, n is fixed), it will allow us for more flexibility when considering asymptotic approximations to the distribution of different statistics.

We implicitly restrict the parameter space of the statistical model in (10) by requiring $\mathbb{P}_n$ to satisfy the following high-level assumption:

Assumption 1 (High-level assumptions on $\mathbb{P}_n$)**.** *The distribution $\mathbb{P}_n$ satisfies*

1. $(1/n)\sum_{i=1}^n x_i x_i^\top \xrightarrow{p} \Sigma$ *for a positive definite matrix* Σ.
2. $(1/\sqrt{n})\sum_{i=1}^n x_i \epsilon_i \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Omega)$ *for a positive definite matrix* Ω.

The high-level Assumption 1 is a convenient way to consider a large class of data generating processes in our analysis. It is well known that Assumption 1 can be verified under i.i.d. sampling and standard primitive conditions on the joint distribution of x_i and ϵ_i (see Assumption 7.2 in Hansen (2022) and Theorems 7.1 and 7.2 therein). However, our framework allows us to incorporate richer data structures, where it is possible to depart from i.i.d. sampling and have general forms of heteroskedasticity and autocorrelation in $(x_i^\top, \epsilon_i)$.

The continuous mapping theorem and Slutsky's theorem immediately imply the asymptotic normality of the least-squares estimator of β_n in the model (10); namely

$$\sqrt{n}\left(\widehat{\beta}_{\text{OLS}} - \beta_n\right) \xrightarrow{d} \mathcal{N}_k\left(\mathbf{0}, \Sigma^{-1}\Omega\Sigma^{-1}\right), \quad \widehat{\beta}_{\text{OLS}} \equiv \left(\sum_{i=1}^n x_i x_i^\top\right)^{-1} \sum_{i=1}^n x_i y_i. \tag{11}$$

This is a slight generalization of the normal approximation to the distribution of the least-squares estimator derived under i.i.d. sampling (for example, see Theorem 7.3 in Hansen (2022)), where, for example, Ω can be the long run variance of the process $\{x_i\epsilon_i\}_{i=1}^\infty$, and where we allow β_n to vary with the sample size.

3 Main Results

3.1 Asymptotic Distribution of the Ridge Estimator

Define the *ridge estimator*—with regularization parameter $\lambda \in \mathbb{R}_+$ and a reference vector $\beta_0 \in \mathbb{R}^k$—to be the solution of the following minimization problem:

$$\min_{\beta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \frac{\lambda}{n} \sum_{j=1}^{k} (\beta_j - \beta_{0j})^2, \tag{12}$$

The solution to the minimization problem in (12) can be shown to equal

$$\widehat{\beta}_\lambda \equiv \left(\sum_{i=1}^{n} x_i x_i^\top + \lambda \mathbb{I}_k \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i + \lambda \beta_0 \right). \tag{13}$$

It is known that the objective function that defines the ridge estimator intentionally sacrifices *training error* (since predictors of y_i based on $\widehat{\beta}_\lambda$ will have a larger training error than predictions based on the least-squares estimator $\widehat{\beta}_{\text{OLS}}$) by penalizing deviations away from the reference vector β_0.

We note that in most textbook definitions of the ridge estimator, the reference vector β_0 equals a vector of zeros of dimension $k \times 1$; see, for example, Chapter 3.4, Equation 3.41 in Friedman et al. (2017) or Chapter 29.5 in Hansen (2022). Since the usual interpretation of (13) is that the ridge estimator "shrinks" the least-squares coefficients, we allow for the possibility that this shrinkage is towards an arbitrary reference vector β_0 that could be different from zero. We note that this more general formulation has several precedents in the literature; for example, Equation 2.1 in Swindel (1976). See also Anup and Maddala (1984).

As we mentioned in the introduction, the main goal of this paper is to present a useful approximation to the finite-sample distribution of the ridge estimator in (13). We are particularly interested in deriving an approximation that captures the well-known fact that, in finite samples, the ridge

estimator presents a bias-variance trade-off. Textbook expressions of the bias and variance of the ridge estimator in finite samples are usually derived assuming i.i.d. sampling and conditional mean independence; i.e., $\mathbb{E}_{\mathbb{P}_n}[\epsilon_i|x_i] = \mathbb{E}_{\mathbb{P}_n}[\epsilon_i] = 0$. See, for example, Chapter 29.6 in Hansen (2022). This stands in contrast with the generality under which the asymptotic distribution of the least-squares estimator can be derived; see Equation (11) above.

Our first result shows that when the regularization parameter λ is *negligible relative to the sample size* (even if it is data dependent), the standard asymptotic distribution of the ridge estimator will not be useful to capture the bias-variance trade-off that is present in finite samples. More precisely, the following result shows that the asymptotic distribution of the ridge estimator will be asymptotically equivalent to that of the least-squares estimator. In order to state our result (and to allow for the possibility that the regularization parameter λ could be selected in a data-driven manner), we consider a possibly stochastic sequence of regularization parameters $\{\lambda_n\}_{n=1}^{\infty}$.

Proposition 1 (Approximation to the distribution of $\widehat{\beta}_{\lambda_n}$ when λ_n/n is negligible)**.** *Suppose that the dataset D_n was generated according to the statistical model* (10)*, and suppose that $\mathbb{P}_n$ satisfies Assumption 1. If*

$$\lambda_n/n \xrightarrow{p} 0 \;\; \textit{and} \;\; (\lambda_n/n)\sqrt{n}(\beta_n - \beta_0) \xrightarrow{p} 0,$$

then

$$\sqrt{n}\left(\widehat{\beta}_{\lambda_n} - \beta_n\right) - \sqrt{n}\left(\widehat{\beta}_{OLS} - \beta_n\right) \xrightarrow{p} 0.$$

Proof. See Appendix A.1. □

The intuition behind this result follows directly from the expression for $\widehat{\beta}_n$ in (13). When n is large and λ_n/n is close to zero, then $\widehat{\beta}_{\lambda_n}$ is approximately the same as $\widehat{\beta}_{\text{OLS}}$. This is why $\sqrt{n}(\widehat{\beta}_{\lambda_n} - \beta_n)$ is approximately the same as the distribution of the least-squares estimator.

The proof of Proposition 1 shows that if λ_n is not negligible relative to the sample size, then the asymptotic distribution of the ridge estimator will differ from that of the least-squares estima-

tor. Moreover, the asymptotic distribution will capture the bias-variance trade-off that the ridge estimator faces in finite samples.

Theorem 1 (Approximation to the distribution of $\widehat{\beta}_{\lambda_n}$ when λ_n/n is non-negligible)**.** *Suppose that the dataset D_n was generated according to the statistical model* (10)*, and suppose that $\mathbb{P}_n$ satisfies Assumption 1. If*

$$\lambda_n/n \xrightarrow{p} \lambda \in [0, \infty) \;\; \textit{and} \;\; \sqrt{n}(\beta_n - \beta_0) \to b \in \mathbb{R}^k, \tag{14}$$

then

$$\sqrt{n}\left(\widehat{\beta}_{\lambda_n} - \beta_n\right) \xrightarrow{d} \mathcal{N}_k\left(-\lambda(\Sigma + \lambda \mathbb{I}_k)^{-1} b, (\Sigma + \lambda \mathbb{I}_k)^{-1} \Omega (\Sigma + \lambda \mathbb{I}_k)^{-1}\right). \tag{15}$$

Proof. See Appendix A.2. □

We view the mean and variance in the limiting distribution in (15) as providing simple generalizations of common formulae for the bias and variance of the ridge estimator derived conditional on covariates and under i.i.d. sampling; see, for example, Equations 29.8 and 29.9 in Hansen (2022).

The key assumptions of Theorem 1 are the two conditions in (14). The first condition requires the regularization parameter to be non-negligible relative to the sample size. Note that we allow for the possibility that the regularization parameter is selected in a data-driven way, as long as it has a deterministic probability limit. We note that the idea of using sequences of regularization parameters λ_n to analyze the asymptotic properties of penalized estimators is not ours and has several precedents in the statistics literature; see, for example, Theorems 2 and 3 in Knight and Fu (2000) where different rate conditions on λ_n are used to analyze the ridge estimator and more general lasso-type estimators.

The second condition in (14) can be interpreted as saying that the true regression coefficient β_n is "local-to-β_0". This means that we are assuming that the reference vector used to compute the ridge estimator in (13) is not too far from the true (and unknown) β_n. We note that without this assumption, the non-negligibility of λ_n leads to an asymptotic distribution where the bias is so

large that β_{λ_n} is no longer $\sqrt{n}$-consistent.

The idea of considering drifting sequences of parameter values in order to improve the distributional approximations provided by standard asymptotic theory has a long history in econometrics. See, for example, the local-to-zero asymptotics in the linear instrumental variables model of Staiger and Stock (1997); the local-to-unit-root asymptotics in the study of nearly integrated autoregressive processes in Phillips (1988) and its recent generalization in Dou and Müller (2021); the local-to-identification-failure analysis in the study of nonlinear Generalized Method of Moments models with weak identification in Andrews and Mikusheva (2022); and the growing-number-of-folds asymptotic framework in Velez (2024), which improves distributional approximations for Debiased Machine Learning estimators. See Powell (2017) for more examples and details on the use of nonstandard asymptotics in econometrics. Some recent work more directly related to our set-up is the paper of Shen and Xiu (2025), studying a high-dimensional linear regression model with *weak* signals. While our work focuses on regression models where the dimension of the covariates is fixed, a version of the weak signals model in Shen and Xiu (2025) can be obtained in the case in which the reference vector equals zero and the true parameter β_n drifts towards zero at rate $1/\sqrt{n}$.

Although Theorem 1 follows from elementary asymptotic theory, we are not aware of previous work that had formally derived the simple approximation in (15) at the level of generality allowed by Assumption 1. Textbook results for the bias and variance of the ridge estimator (usually derived conditional on covariates and under i.i.d. sampling of a linear regression model) can yield a Gaussian distribution similar to (15) in finite samples if the error term in the regression model is also assumed to be Gaussian (conditional on the vector of covariates). But the analysis is more nuanced if the error distribution is non-Gaussian and/or there is serial autocorrelation in the regression residuals. The closest expression to (15) that we were able to find in the literature is given in the seminal paper of Knight and Fu (2000) providing asymptotics for lasso-type estimators (including the ridge estimator). On p. 1368 they use their Theorem 4 to derive an approximation to the asymptotic distribution of a ridge estimator with reference vector $\beta_0 = 0_{k\times 1}$. Under i.i.d. sampling and a

conditional homoskedasticity assumption (i.e., $\Omega = \mathbb{E}[x_i x_i^\top \epsilon_i^2] = \mathbb{E}[x_i x_i^\top]\mathbb{E}[\epsilon_i^2]$), they obtain the same formula as in (15) but specialized to the case in which $\Omega = \mathbb{E}[\epsilon_i^2]\Sigma$. Thus, our results can be viewed as a generalization of their analysis of the asymptotic distribution of the ridge estimator, but allowing for more general data structures (where heteroskedasticity and serial autocorrelation are permitted and captured by the matrix Ω), and where we also allow for a general reference vector β_0. Knight and Fu (2000) also consider an asymptotic sequence where β_n changes with the sample size (they refer to this as *triangular array asymptotics*). We note that a potentially interesting extension of our results is to derive asymptotic approximations for lasso-type estimators (analogous to part b) of their Theorem 4) under our Assumption 1. These results could then be used to provide concrete recommendations for the choice of regularization parameter. In the remaining part of the paper, we maintain our focus on the ridge estimator and use the asymptotic approximation in Theorem 1 to present approximations to the *prediction risk* of the ridge estimator, and we use statistical decision theory to provide recommendations on the choice of regularization parameter.

3.2 An approximation to the Prediction Risk of Ridge Regression

In this section we show that it is possible to use Theorem 1 to provide an approximation to the *prediction risk* of ridge regression. We start by providing a definition of the prediction problem we are interested in, and we present a formal definition of the risk of predictions based on ridge regression.

Prediction Problem. An econometrician has access to a dataset $D_n \equiv \{(y_i, x_i^\top)\}_{i=1}^n$ comprised of n observations of a real-valued outcome variable, $y_i \in \mathbb{R}$, and a vector of k covariates, $x_i \in \mathbb{R}^k$. The dataset D_n is assumed to have been generated by the statistical model (10) with parameters $(\beta_n, \mathbb{P}_n)$ that satisfy Assumption 1. A prediction function is a mapping $a : \mathbb{R}^k \to \mathbb{R}$ that maps covariates into a predicted value for the outcome variable. We note that a prediction function is defined for all possible values of covariates, including those that have not been observed in the sample. In order to define a loss function for the prediction problem, we introduce an additional assumption.

Assumption 2. *The stochastic process* $\{(x_i^\top, \epsilon_i)\}_{i=1}^\infty$ *is strictly stationary in the sense of Definition 1.3.3 in Brockwell and Davis (2013).*

The strict stationarity assumption means that—associated to $\mathbb{P}_n$—there exists a probability distribution $\mathbb{P}$ over $\mathbb{R}^{k+1}$ such that, for every $i = 1, \ldots, n$, we have $(x_i^\top, \epsilon_i) \overset{\mathbb{P}_n}{\sim} \mathbb{P}$. We will refer to $\mathbb{P}$ as the stationary distribution of $\mathbb{P}_n$. We use this stationary distribution to define a loss function for the prediction problem as follows. Suppose we sample a new pair $(x^\top, \epsilon)$ according to the stationary distribution $\mathbb{P}$, and use the slope coefficient β_n to construct a new outcome-covariate pair $(y, x) = (x^\top \beta_n + \epsilon, x)$.[3] We can then define the loss associated to a prediction function $a : \mathbb{R}^k \to \mathbb{R}$ as

$$L(a, \beta_n, \mathbb{P}) \equiv \mathbb{E}_{(y,x)\sim(\beta_n,\mathbb{P})}[(y - a(x))^2].$$

In a slight abuse of notation, $(y, x) \sim (\beta_n, \mathbb{P})$ is used to capture the fact that the expectation is taken over outcome-covariate pairs generated using $(\beta_n, \mathbb{P})$ as described above.

Prediction Risk. In a prediction problem, the goal is to construct a data-driven prediction function; that is, we want to use the available training data to choose a prediction function. In a slight abuse of notation, we denote data-driven prediction functions as $\widehat{a}$. Formally, we think of $\widehat{a}$ as a mapping that takes the data set D_n as input and returns a prediction function $\widehat{a}$ from covariates to outcomes. Because $\widehat{a}$ is constructed from the training data, it is a random function. The prediction risk is then the expected prediction loss

$$R(\hat{a}; \beta_n, \mathbb{P}_n) = \mathbb{E}_{(\beta_n,\mathbb{P}_n)}[\mathcal{L}(\widehat{a}; \beta_n, \mathbb{P})], \tag{16}$$

where $\mathbb{E}_{(\beta_n,\mathbb{P}_n)}$ means that the expectation is taken over the distribution of D_n, which is parameter-

[3]Note then that in our analysis we do not allow the distribution used to evaluate prediction error to deviate arbitrarily from the distribution that generated the data. As shown by Patil, Du, and Tibshirani (2024), such an assumption can have important implications over the optimal choice of regularization parameter. A framework for analyzing the distributionally robust prediction error for the square-root lasso and related estimators (and for optimally selecting tuning parameters) has been proposed in Montiel Olea, Rush, Velez, and Wiesel (2026), but their framework excludes the ridge regression estimator.

ized by $(\beta_n, \mathbb{P}_n)$.

Approximate Prediction Risk. Consider the prediction function based on the ridge regression estimator

$$\widehat{a}_{\lambda_n}(x) = x^\top \widehat{\beta}_{\lambda_n}, \tag{17}$$

where $\widehat{\beta}_\lambda$ is defined in (13).

The exact finite-sample analysis of the prediction risk of ridge regression is challenging even under the stylized homoskedastic, Gaussian regression model. However, there are different papers that provide approximations (and lower/upper bounds) to this prediction risk. For example, Dobriban and Wager (2018), Hastie et al. (2022), Mourtada and Rosasco (2022), Atanasov et al. (2024). The following result shows that we can use Theorem 1 to provide an approximation to the excess prediction risk of ridge regression as a function of three types of parameters: λ (the probability limit of λ_n/n); b (the parameter controlling the local-to-β_0 regression coefficient β_n); and the variance parameters Σ, Ω (where $\Sigma = \mathbb{E}_{\mathbb{P}}[xx^\top]$, and Ω is the asymptotic variance defined in Assumption 1). More precisely, define the *approximate excess risk function*

$$R_n^e(\lambda; b, \Sigma, \Omega) \equiv \frac{1}{n}\lambda^2 b^\top (\Sigma + \lambda \mathbb{I}_k)^{-1}\Sigma(\Sigma + \lambda \mathbb{I}_k)^{-1} b + \frac{1}{n}\text{trace}\left((\Sigma + \lambda \mathbb{I}_k)^{-1}\Omega(\Sigma + \lambda \mathbb{I}_k)^{-1}\Sigma\right). \tag{18}$$

Theorem 2 (Approximation to the excess prediction risk of ridge regression)**.** *Suppose that the dataset D_n was generated according to the statistical model* (10)*, and suppose that $\mathbb{P}_n$ satisfies Assumptions 1 and 2. If*

(i) $\lambda_n/n \xrightarrow{p} \lambda \in [0, \infty)$ *and* $\sqrt{n}(\beta_n - \beta_0) \to b \in \mathbb{R}^k$*,*

(ii) there exist estimators $(\widehat{\Sigma}, \widehat{\Omega})$ *that are consistent for* (Σ, Ω)*,*

(iii) there exists $\delta > 0$ *such that*

$$\sup_n \mathbb{E}_{\mathbb{P}_n}\left[(Z_n^\top \Sigma Z_n)^{1+\delta}\right] < \infty, \qquad Z_n \equiv \sqrt{n}(\widehat{\beta}_{\lambda_n} - \beta_n),$$

then

$$R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2 = R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega}) + o_{(\beta_n, \mathbb{P}_n)}(1/n), \tag{19}$$

where $\sigma^2 \equiv \mathbb{E}_{\mathbb{P}}[\epsilon^2]$.

Proof. See Appendix A.3 □

The difference between the finite sample prediction risk of a given data-driven predictor $\widehat{a}$—which we have denoted by $R_n(a_n; \beta_n, \mathbb{P}_n)$—and the residual variance σ^2 is typically referred to as the *excess prediction risk* or simply *excess risk*.[4] Theorem 2 says that, in large samples, the excess prediction risk of ridge regression equals—up to a term that is small in probability—by the approximate excess prediction risk function defined in Equation (18), evaluated at $(\lambda, b, \widehat{\Sigma}, \widehat{\Omega})$. The result shows that if we fix $(\widehat{\Sigma}, \widehat{\Omega})$, the excess risk is expected to vary as a function of i) how large is the penalty parameter relative to the sample size (λ is the probability limit of λ_n/n) and ii) how close is the reference vector β_0 to the true coefficient β_n (b is the limit of $\sqrt{n}(\beta_n - \beta_0)$).

Theorem 2 is conceptually related to an active area of research in the statistics literature providing asymptotic and nonasymptotic approximations to the excess prediction risk of the ridge regression estimator; see, for example, Dobriban and Wager (2018),Hastie et al. (2022), Mourtada and Rosasco (2022) and the references therein. All these papers make distributional assumptions that preclude data generating processes that exhibit conditional heteroskedasticity, time-series autocorrelation, or a nonlinear conditional expectation function. The additional assumptions in these papers, however, allow them to consider *high-dimensional* asymptotics where the number of covariates is allowed to grow relative to the sample size. In contrast, all the analysis in this paper pertains a *low-dimensional* model where we treat the number of covariates as fixed, and let the sample size diverge to infinity.

[4] See, for example, Mourtada (2022).

3.3 Selection of λ

In this section, we use the approximation to the excess prediction risk of ridge regression given in Theorem 2 to make a concrete recommendation regarding the selection of the regularization parameter λ_n. The key insight is to pick λ—which is the probability limit of λ_n/n—to optimize the function $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$, which is the leading term in (19). The main conceptual challenge is that approximation depends on the unknown localization parameter $b \in \mathbb{R}^k$, which was defined to be the limit of the sequence $\sqrt{n}(\beta_n - \beta_0)$. Since b cannot be estimated consistently, we suggest to handle this *localization* parameter in a way that we think is entirely analogous to the evaluation of risk functions in statistical decision theory: optimizing either the worst-case or the average risk. Our suggested method provides a well-defined mapping from the data (through the estimated covariance matrices $\widehat{\Sigma}$ and $\widehat{\Omega}$) to values of the regularization parameter. We provide details below.

3.3.1 Choosing λ to minimize the worst-case approximate excess risk

We first consider the problem of minimizing the worst-case approximate excess risk. Let $\mathcal{B} \subseteq \mathbb{R}^k$ be a user-specified set of potential values for b. For example, one could take $\mathcal{B} \equiv \{b \in \mathbb{R}^k \quad \|b\| \leq B\}$. for some known $B > 0$, where $\|\cdot\|$ denotes the standard Euclidean norm. We say that $\widehat{\lambda}_n^*$ is a $\mathcal{B}$-minimax selection of the regularization parameter in ridge regression if $\widehat{\lambda}_n^* = n \cdot \widehat{\lambda}^*$ where $\widehat{\lambda}^*$ solves the minimax problem

$$\inf_{\lambda \geq 0} \sup_{b \in \mathcal{B}} R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega}). \tag{20}$$

We make two comments about the minimax problem in (20). First, if the set $\mathcal{B}$ is unbounded, then the $\mathcal{B}$-minimax choice of the regularization parameter in ridge regression is $\widehat{\lambda}_n = 0$. This means that in order to obtain a different choice of $\widehat{\lambda}_n$ under the minimax criterion we will require a bound on the parameter $b \in \mathbb{R}^k$.

Second, while the outer optimization problem in (20) could be handled via grid search (as it is done whenever $\widehat{\lambda}_n$ is chosen by means of cross-validation), the inner optimization problem requires

the evaluation of the worst-case approximate excess risk. Since only the first term in the expression of $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ in Equation (18) depends on b, we can solve for the worst-case risk by solving the following constrained optimization problem:

$$\widehat{v}_{\text{worst-case}}(\lambda) \equiv \sup_b b^\top (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} b \quad \text{subject to } b \in \mathcal{B}. \tag{21}$$

Algebra shows that when $\mathcal{B} \equiv \{b \in \mathbb{R}^k \ \ \|b\| \leq B\}$—for some known $B > 0$, and with $\|\cdot\|$ given by the Euclidean norm—the constrained optimization problem in (21) can be solved in closed-form up to a maximum eigenvalue computation:

$$\widehat{v}_{\text{worst-case}}(\lambda) = B^2 \cdot \text{max-eigenvalue} \left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \right).$$

Thus, in the special case in which the Euclidean norm of b is bounded by B, the $\mathcal{B}$-minimax choice of $\widehat{\lambda}_n$ can be implemented as $\widehat{\lambda}_n^* = n \times \widehat{\lambda}^*$ where $\widehat{\lambda}^*$ solves the problem

$$\inf_{\lambda \in \mathbb{R}_+} \lambda^2 B^2 \text{max-eigenvalue} \left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \right) + \text{trace} \left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Omega} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} \right). \tag{22}$$

The minimization problem in (22) is a nonlinear (and nonconvex) optimization problem over $\mathbb{R}_+$ and, in general, does not have a closed-form solution. Our suggestion is to minimize this function over the same grid of parameter values used in the standard implementation of (leave-one-out) cross-validation for ridge regression provided in standard statistical packages such as `glmnet`.

3.3.2 Choosing λ to minimize average approximate excess risk

We now consider the problem of choosing the regularization parameter of ridge regression by minimizing the *average* approximate excess risk. Let π denote a user-specified probability distribution over $\mathbb{R}^k$. We say that $\widehat{\lambda}_n^*$ is a π-optimal selection of the regularization parameter in ridge regression

if $\widehat{\lambda}_n^* = n \cdot \widehat{\lambda}^*$ where $\widehat{\lambda}^*$ solves the problem

$$\inf_{\lambda \geq 0} \mathbb{E}_{b \sim \pi}[R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})]. \tag{23}$$

Once again, since only the first term in the expression of $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ in Equation (18) depends on b, we can solve for the average excess risk by evaluating the following expectation:

$$\widehat{v}_{\text{average}}(\lambda) \equiv \mathbb{E}_{b \sim \pi}\left[b^\top (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} b\right]. \tag{24}$$

This expectation can be evaluated analytically in some cases. For example, suppose $b \sim \mathcal{N}_k\left(\mathbf{0}, C^2 \mathbb{I}_k\right)$, where C is a user-specified hyper-parameter. Algebra shows that

$$\widehat{v}_{\text{average}}(\lambda) = C^2 \cdot \text{trace}\left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1}\right).$$

A similar formula can be derived if the distribution over b is elliptical with mean zero and a covariance matrix proportional to the identity matrix; see Appendix B.1. The π-optimal choice of $\widehat{\lambda}_n$ in these cases is $\widehat{\lambda}_n^* = n \times \widehat{\lambda}^*$, where $\widehat{\lambda}^*$ solves the problem

$$\inf_{\lambda \in \mathbb{R}_+} \lambda^2 C^2 \text{trace}\left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1}\right) + \text{trace}\left((\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Omega} (\widehat{\Sigma} + \lambda \mathbb{I}_k)^{-1} \widehat{\Sigma}\right). \tag{25}$$

Just as for the $\mathcal{B}$-minimax choice of regularization parameter, our suggestion is to minimize this function over the same grid of parameter values used in the standard implementation of (leave-one-out) cross-validation for ridge regression provided in standard statistical packages.

3.3.3 Isotropic Features

We now specialize the derivations in Sections 3.3.1 and 3.3.2 to the case in which Σ—the matrix of second moments of the vector of covariates—is known to be proportional to identity matrix; that is,

$\Sigma = \sigma_x^2 \mathbb{I}_k$. Hastie et al. (2022) refer to predictions problems of this form as prediction with *isotropic features*.[5] Algebra shows that with isotropic features, both the minimax and average-risk objectives outlined in the previous subsections simplify substantially, allowing explicit characterization of the optimal regularization parameter for ridge regression. The key observation is that under isotropic features, the approximate excess risk in Equation (18) simplifies to

$$R_n^e(\lambda; b, \sigma_x^2, \Omega) = \frac{1}{n} \frac{\lambda^2 \sigma_x^2}{(\sigma_x^2 + \lambda)^2} b^\top b + \frac{1}{n} \frac{\sigma_x^2}{(\sigma_x^2 + \lambda)^2} \operatorname{trace}(\Omega). \tag{26}$$

The following theorem presents the optimal selection of the regularization parameter for ridge regression under the minimax and average risk criterion.

Theorem 3 (Optimal choice of regularization parameter with isotropic features)**.** *Suppose that the assumptions of Theorem 2 hold. Assume that* $\Sigma = \sigma_x^2 \mathbb{I}_k$, $\sigma_x^2 > 0$, *and that there exist estimators* $(\widehat{\sigma}_x^2, \widehat{\Omega})$ *that are consistent for* (σ_x^2, Ω). *Then:*

(1) If $\mathcal{B} \equiv \{b \in \mathbb{R}^k \quad \|b\| \leq B\}$*—for some known* $B > 0$*, and with* $\|\cdot\|$ *given by the Euclidean norm—the* $\mathcal{B}$*-minimax selection of the regularization parameter in ridge regression is*

$$\widehat{\lambda}^*_{\text{minimax}} = n \cdot \frac{\operatorname{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 B^2}.$$

(2) If $\mathbb{E}_\pi[b^\top b] < \infty$ *and* $\mathbb{E}_\pi[b^\top b] > 0$*, the* π*-optimal selection of the regularization parameter in ridge regression is*

$$\widehat{\lambda}^*_\pi = n \cdot \frac{\operatorname{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 \mathbb{E}_\pi[b^\top b]}.$$

Proof. See Appendix A.4 □

Since the introduction of the ridge regression estimator, the problem of providing a concrete recommendation for its regularization parameter has been analyzed in multiple papers; for example,

[5] Hastie et al. (2022) further assume that $\sigma_x^2 = 1$.

see Gibbons (1981) and the references therein. The idea of using an approximation of the risk function to choose regularization parameters has a long precedent in statistics and econometrics; see, for example the discussion of Abadie and Kasy (2019) on p. 744. To the best of our knowledge the formulae we provide in Theorem 3 are new to the literature. For the formula we provide in (2), the closest reference we were able to find in the literature is the optimal regularization parameter derived in Theorem 2.1 in Dobriban and Wager (2018), p. 254. Their recommendation is based on high-dimensional model where covariates and regression residuals are independent, but covariates need not be isotropic. Assuming a prior on regression coefficients such that $\mathbb{E}_\pi[\beta] = 0$ and $\mathrm{Var}_\pi[\beta] = \alpha^2 \mathbb{I}_k/k$ (see their Assumption RRC), Dobriban and Wager (2018) recommend a regularization parameter of the form

$$\hat{\lambda} = n \cdot \frac{k}{n} \frac{1}{\alpha^2} = \frac{k}{\alpha^2}.$$

In the homoskedastic case with isotropic covariates—$\Omega = \sigma_\epsilon^2 \mathbb{I}_k$—our formula matches their recommendation since

$$\widehat{\lambda}^*_\pi = n \cdot \frac{\mathrm{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 \mathbb{E}_\pi[b^\top b]} = \frac{k}{\mathbb{E}_\pi[\beta_n^\top \beta_n]} = \frac{k}{\alpha^2}.$$

However, our result allows for heteroskedasticity and autocorrelation in the data generating process by adjusting the regularization parameter as a function of Ω. The main limitation of our result, as discussed before, is that our theory is only applicable to low-dimensional models.

More generally, it is important to mention again that one of the initial motivations of our paper was to recover an expression analogous to the optimal regularization parameter for the ridge regression estimator given by Hastie et al. (2022) in their Corollary 6 to Theorem 6. Their results show that using high-dimensional asymptotics where $k/n \to \gamma$, the oracle choice of regularization parameter for the ridge estimator (assuming x_i and ϵ_i are independent and that $\Sigma = \mathbb{I}_k$) takes the form $\lambda^* = n\mathbb{E}[\epsilon_i^2]\gamma/\|\beta\|^2$. Assuming conditional homoskedasticity, our formulae in (8) and (9) become analogous to Hastie et al. (2022)'s formula, but the true unknown β_n is replaced by either

the prior mean of $\|b\|^2$ or the maximum value of $\|b\|^2$.

4 Simulations

We now examine the extent to which the asymptotic approximations presented in Section 3 capture the finite-sample behavior of the ridge estimator. For this purpose, we present Monte-Carlo simulations to calculate and compare the prediction risk of the following estimators:

1. The ridge regression with the regularization parameter chosen to minimize the worst-case approximate excess-risk criterion in (22).

2. The ridge estimator tuned according to the standard leave-one-out cross-validation; see Section 3 in Patil, Wei, Rinaldo, and Tibshirani (2021).

3. The OLS estimator.

Henceforth, we refer to these estimators as minimax ridge, LOO-CV ridge, and OLS, respectively.

4.1 DGP-1

We first consider the stylized Gaussian homoskedastic linear regression model

$$y_i = x_i^\top \beta_n + \epsilon_i, \qquad x_i \sim N(0, \sigma_x^2 I_k), \qquad \epsilon_i \sim N(0, \sigma^2), \qquad \beta_n = \beta_0 + \frac{b}{\sqrt{n}},$$

with $\beta_0 = (0, \ldots, 0)^\top$. We assume that ϵ_i is independent of x_i. Let $\Sigma = \mathbb{E}[x_i x_i^\top] = \sigma_x^2 I_k$, and $\Omega = \mathbb{E}[x_i x_i^\top \epsilon_i^2]$. Under homoskedasticity and independence, $\Omega = \sigma^2 \Sigma = \sigma^2 \sigma_x^2 I_k$.

Throughout this subsection we set $k = 10$, $\sigma_x^2 = 1$, $\sigma^2 = 1$, and $b = (1, \ldots, 1)^\top$. We set $B \equiv \|b\| = \sqrt{10}$. In our simulations, we consider six different sample sizes: $n \in \{500, 1000, 1500, 2000, 2500, 3000\}$. Each simulation uses 2,000 Monte Carlo repetitions.

We use the term *minimax ridge* to refer to the ridge estimator that uses the regularization parameter given by (27) and β_0 as reference vector. Since the vector of covariates has isotropic features ($\Sigma = \sigma_x^2 I_k$), the regularization parameter is chosen as in Theorem 3; that is:

$$\widehat{\lambda}_{\text{minimax}} = n \frac{\text{trace}(\widehat{\Omega})}{\widehat{\sigma}_x^2 B^2}, \tag{27}$$

where $\widehat{\sigma}_x^2$ is estimated from the sample covariance matrix of the covariates, and Ω is consistently estimated by the feasible sample analogue of $\mathbb{E}[x_i x_i^\top \epsilon_i^2]$, obtained by replacing ϵ_i with OLS residuals (and imposing conditional homoskedasticity).

For the LOO-CV (leave-one-out cross-validation) ridge, we use the leave-one-out cross-validation procedure as in Patil et al. (2021, Section 3).[6] The candidate set of regularization parameters $\lambda_n = nc$ is based on a nonuniform grid of 500 strictly positive candidate values of c spanning $[0.001, 5]$. We let `RidgeCV` choose the minimizer of the leave-one-out criterion over that grid.

For each sample size n and $m = 1, \ldots, 2000$, we proceed as follows:

1. Draw an independent sample $D_n^{(m)} = \{(Y_i^{(m)}, X_i^{(m)})\}_{i=1}^n$ from the Gaussian model.
2. Calculate the minimax ridge, LOO-CV ridge, and OLS using $D_n^{(m)}$.
3. Calculate the prediction error for each estimator $\widehat{\beta}_n^{(m)}$,

$$\mathbb{E}_{(Y,X)\sim(\beta_n,\mathbb{P})}\left[(Y - X^\top \widehat{\beta}_n^{(m)})^2\right] = \sigma^2 + (\widehat{\beta}_n^{(m)} - \beta_n)^\top \Sigma (\widehat{\beta}_n^{(m)} - \beta_n).$$

We then approximate the prediction risk by averaging the prediction errors across repetitions:

$$\frac{1}{2000} \sum_{m=1}^{2000} \mathbb{E}_{(Y,X)\sim(\beta_n,\mathbb{P})}\left[(Y - X^\top \widehat{\beta}_n^{(m)})^2\right].$$

[6] We set `sklearn.linear_model.RidgeCV` with `cv=None` and `fit_intercept=False`.

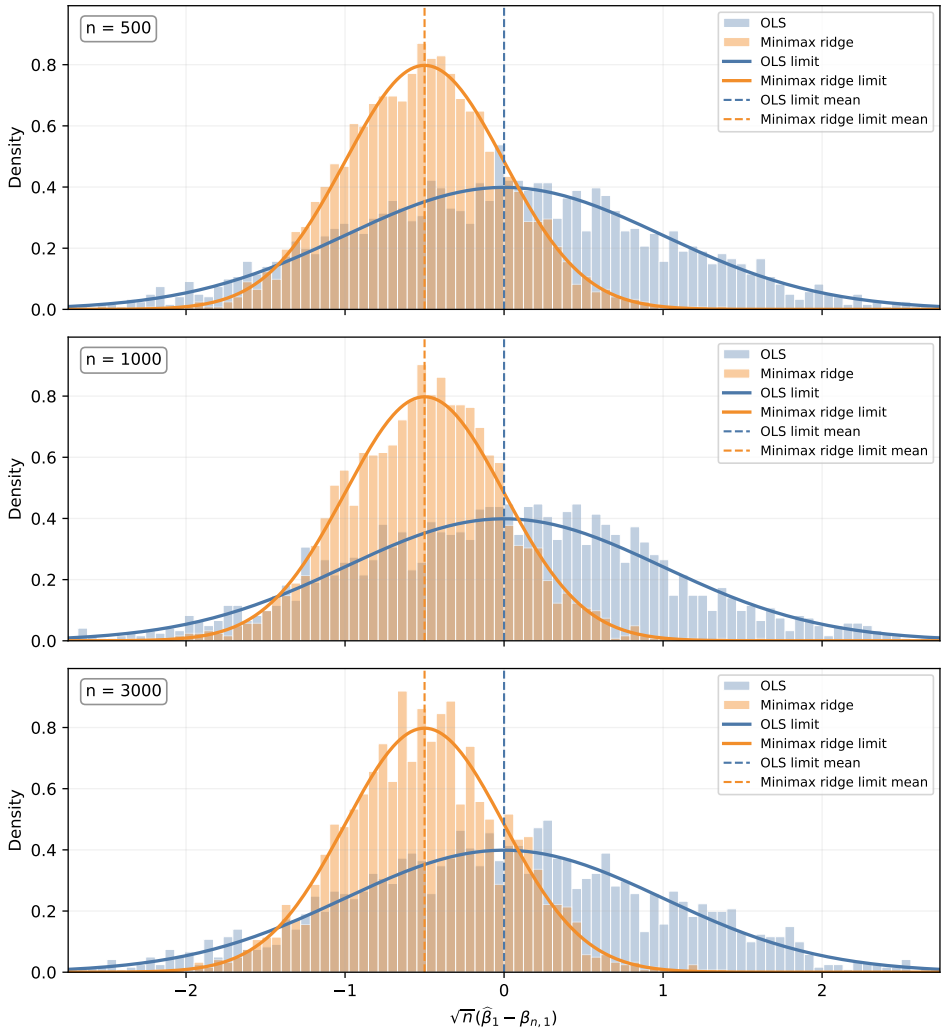

Figure 1: Distribution of the first coefficient under DGP-1

Figure 1 reports the simulated distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_{n,1})$ for OLS and minimax ridge, with the Gaussian limits implied by Theorem 1: $N(0, \frac{\sigma^2}{\sigma_x^2})$ for OLS and $N(-\lambda b_1/(\sigma_x^2 + \lambda), \frac{\sigma^2\sigma_x^2}{(\sigma_x^2+\lambda)^2})$ for minimax ridge. The approximation is accurate already at $n = 500$ and remains so as n grows. Consistent with our asymptotic approximations, the minimax ridge distribution is shifted away from the origin, but markedly more concentrated than the OLS distribution. In contrast, OLS is correctly centered but presents a large dispersion.

We now compare the estimators using the excess prediction risk; that is, $R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2$.

Figure 2 compares the excess prediction risk of the three estimators relative to minimax ridge estimator across different sample sizes. The figure suggests that, in this design, our recommended minimax ridge outperforms both LOO-CV ridge and OLS. Across sample sizes, the risk of LOO-CV ridge is about 20 percent higher than that of minimax ridge.

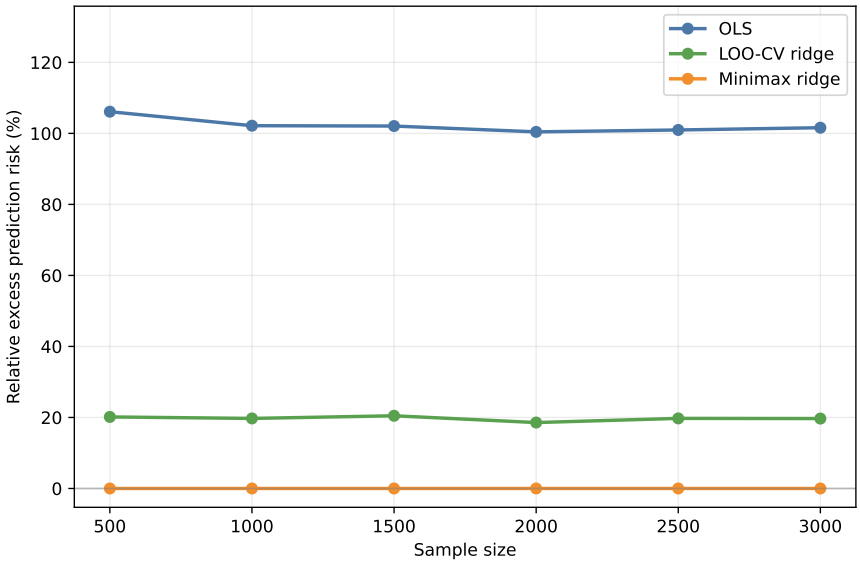

Figure 2: Relative excess prediction risk under DGP-1

We next evaluate our approximation to the excess prediction risk of ridge estimators presented in Theorem 2. Figure 3 displays a comparison between the Monte Carlo estimate of excess prediction risk $R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2$ (solid orange line) and our *feasible* approximate excess-risk $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ (shaded region) for several values of the scaled regularization parameter $\lambda/n \in [0, 5]$.[7] In all the panels, we scale the vertical axis by the sample size n and report results consistent with Equation (19), showing that the approximate excess-risk calculation tracks the excess-risk curve closely. Figure 3 also presents the approximate excess risk minimizer (vertical dashed line), which is the median of the penalized parameter based on Equation 27. For all three sample sizes (n = 500, 1,000, and 3,000), the median selected tuning ratio is approximately 0.99, which is close to the theoretical

[7]For each value of λ, the approximate excess-risk $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ is data-dependent; therefore, for each simulation we have a different value. To capture its range of values, we report the central 95 percent range after dropping the lowest 2.5 percent and highest 2.5 percent of values for each candidate ratio.

value $\lambda_n/n = 1$.

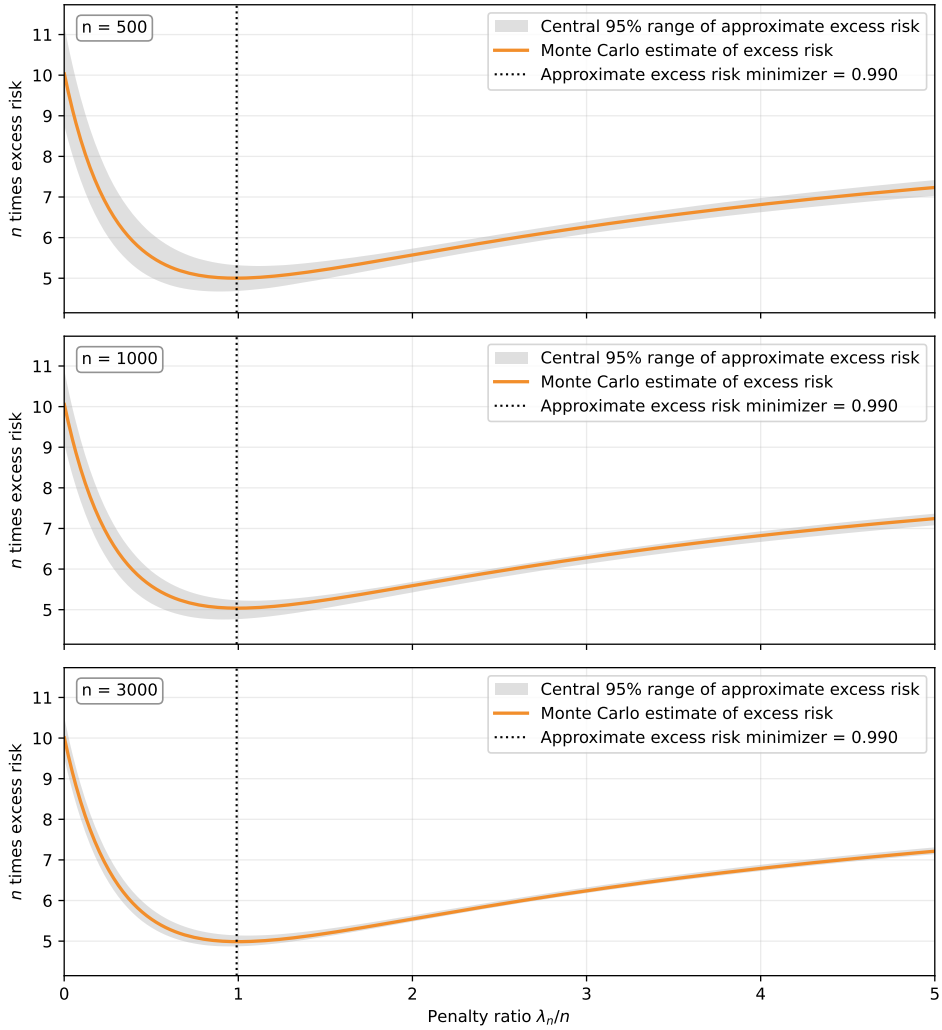

Figure 3: Risk curves under DGP-1

We also compare the scaled regularization parameter ($\widehat{\lambda}_n/n$) used by the LOO-CV and minimax ridge estimators. Figure 10 in Appendix B.2 confirms that the scaled regularization parameter of the minimax ridge—based on (27)—is close to its limit $\lambda_n/n = 1$, as expected. By contrast, the scaled regularization parameter of the LOO-CV ridge is widely dispersed across the values it can take.

4.2 DGP-2

We next consider a low-dimensional version of the weak-signal design based on Section 3.1 of Shen and Xiu (2025). Our asymptotic theory focused on models in which the number of covariates is small relative to the sample size. Thus, before considering a high-dimensional version of this design, it is useful to first understand the performance of our approximations in a non-isotropic design with a smaller number of covariates. We note that this design (and also the one presented in the next section) imply a minor departure from Assumption 2 (strict stationarity) by allowing heterogenous variances.

Following Shen and Xiu (2025), we let the outcome be generated by the linear model $y_i = x_i^\top \beta_n + \epsilon_i$, with $n = 500$ and $k = 50$. The ridge reference vector is $\beta_0 = 0_k$, so the local parameter is $b = \sqrt{n}(\beta_n - \beta_0) = \sqrt{n}\beta_n$. The design matrix is generated as $X = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$, where $Z \in \mathbb{R}^{n\times k}$ has i.i.d. standard normal entries, $\Sigma_1 \in \mathbb{R}^{n\times n}$ controls the dependence across observations, and $\Sigma_2 \in \mathbb{R}^{k\times k}$ controls the dependence across regressors. For Σ_1, we follow Shen and Xiu (2025) and set $(\Sigma_1)_{ij} = 2^{-|i-j|}$ for $1 \le i, j \le n$. For Σ_2, we also follow Shen and Xiu (2025) and construct the second moments of covariates by orthogonal diagonalization so that $\Sigma_2 = Q\Lambda Q^\top$. The orthogonal matrix Q is randomly drawn once, the eigenvalues in Λ are drawn once from $U[0.5, 1.5]$, and Σ_2 is then held fixed throughout the simulation. We note that, by construction, the matrix Σ_2 is not isotropic. In Appendix B.2.2 we describe the distribution of eigenvalues of this matrix and the coefficient vector used in the simulation. Since the diagonal entries of Σ_1 are equal to one, we can show that the population prediction covariance is $\Sigma = \mathbb{E}[x_i x_i^\top] = \Sigma_2$. The errors are independent of the regressors and are generated with diagonal heteroskedasticity, with the diagonal entries of the error-variance matrix drawn once from $U[0.5, 1.5]$ and held fixed. Thus, the asymptotic variance Ω in Assumption 1 is $\Omega = \bar{\sigma}_\epsilon^2 \Sigma_2$, where $\bar{\sigma}_\epsilon^2 \equiv n^{-1} \sum_{i=1}^n \bar{\sigma}_{\epsilon,i}^2$, which is close to one in this design and equals one asymptotically by the strong law of large numbers.

Following the coefficient-generation model in Shen and Xiu (2025, Section 3.1), we first draw a

preliminary coefficient vector $\widetilde{\beta}$, with independent coordinates satisfying

$$\widetilde{\beta}_j \sim 0.2 \cdot 0 + 0.8 \cdot N(0, 1.25), \qquad j = 1, \ldots, k.$$

For each target value of R^2 (5%, 20%, and 50%), we rescale this same draw to obtain the corresponding true coefficient vector β_n. The resulting β_n is then held fixed across all Monte Carlo repetitions.

For the minimax criterion, we calibrate the radius B from the same coefficient-generation scheme: for each target R^2, we draw 2,000 coefficient vectors, rescale each draw to the target R^2, compute $\sqrt{n}\|\beta_n\|$, and use the 90th percentile as the baseline value of B. The $\sqrt{n}$ normalization is the one used in the local parameter $b = \sqrt{n}\beta_n$. To accommodate the finite-sample effect of estimating a non-negligible number of coefficients, we adjust the approximate excess risk function by multiplying the variance term by $n/(n-k)$. Since this design is not isotropic, minimax ridge uses the general non-isotropic approximate excess-risk criterion implied by Theorem 2. In each Monte Carlo repetition, we draw a new sample $(X^{(m)}, Y^{(m)})$, and we plug in the sample analogues $\widehat{\Sigma}_{\text{df}}^{(m)} = X^{(m)\top}X^{(m)}/(n-k)$, $\widehat{\sigma}_{\epsilon,\text{df}}^{2,(m)} = \frac{1}{n-k}\sum_{i=1}^{n} \widehat{u}_{i,\text{OLS}}^{(m)2}$, and $\widehat{\Omega}_{\text{df}}^{(m)} = \widehat{\sigma}_{\epsilon,\text{df}}^{2,(m)}\widehat{\Sigma}_{\text{df}}^{(m)}$. For each target R^2 and each Monte Carlo repetition, we follow the same steps as in DGP-1. The number of Monte Carlo repetitions is 2,000.

In each Monte Carlo repetition m, over a grid of candidate values $\lambda = \lambda_n/(n-k)$, we choose $\widehat{\lambda}_{\text{minimax}}^{(m)}$ to minimize the sum of

$$\lambda^2 B^2 \,\text{maxeigenvalue}\left[(\widehat{\Sigma}_{\text{df}}^{(m)} + \lambda I_k)^{-1}\widehat{\Sigma}_{\text{df}}^{(m)}(\widehat{\Sigma}_{\text{df}}^{(m)} + \lambda I_k)^{-1}\right] \tag{28}$$

and

$$\frac{n}{n-k}\,\text{trace}\left[(\widehat{\Sigma}_{\text{df}}^{(m)} + \lambda I_k)^{-1}\widehat{\Omega}_{\text{df}}^{(m)}(\widehat{\Sigma}_{\text{df}}^{(m)} + \lambda I_k)^{-1}\widehat{\Sigma}_{\text{df}}^{(m)}\right]. \tag{29}$$

Relative to the criterion in Equation (22), the only modification consists of replacing the plug-in covariance matrices by their degrees-of-freedom adjusted analogues and multiplying the variance

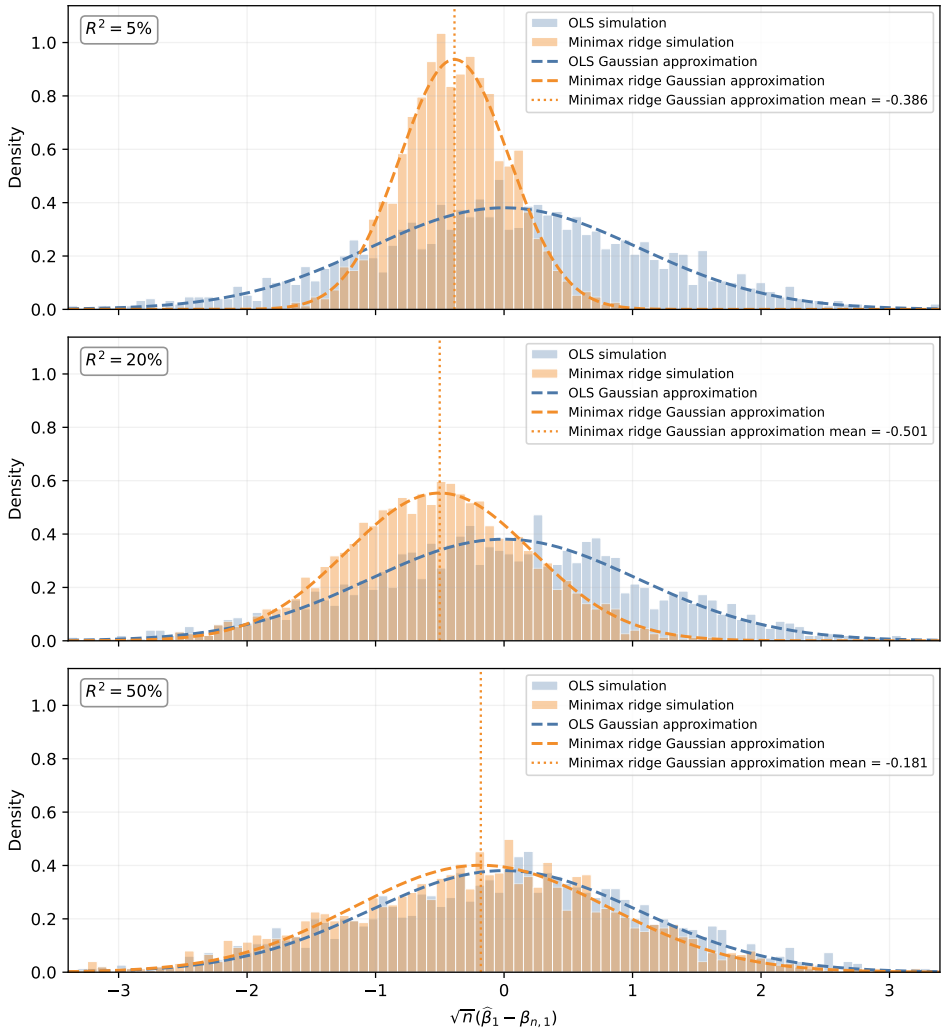

Figure 4: Distribution of the first coefficient under DGP-2

component by $n/(n-k)$. After the minimization, the reported ridge tuning parameter is $\widehat{\lambda}_n^{(m)} = (n-k)\widehat{\lambda}_{\text{minimax}}^{(m)}$, so all selected tuning ratios are reported on the scale $\widehat{\lambda}_n/n$.

For the LOO-CV ridge estimator, we use the same implementation as in DGP-1.[8] The number of Monte Carlo repetitions is 2,000 for each target value of R^2.

The following figures report our simulation results. Figure 4 reports the simulated distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_{n,1})$ for OLS and minimax ridge, together with the Gaussian approximations implied

[8] `sklearn.linear_model.RidgeCV` with `cv=None` and `fit_intercept=False`.

by our theory. The Gaussian curves are computed using the population matrices in the simulation, rather than fitted to the Monte Carlo histograms. Similar to the pattern observed in DGP-1, comparing with the OLS results, the distribution of the minimax ridge estimator is slightly biased but has a smaller variance.

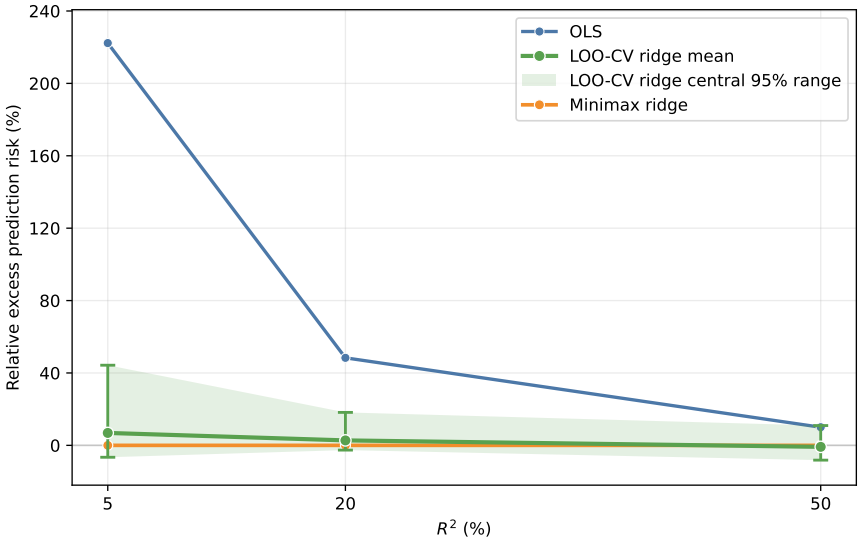

Figure 5: Relative excess prediction risk under DGP-2, LOO-CV and minimax

We next compare the three estimators using excess prediction risk. Figure 5 reports the excess prediction risk of OLS and LOO-CV ridge relative to minimax ridge. The shaded region reports the central 95 percent range of the repetition-specific relative excess risk of LOO-CV ridge. The figure shows that OLS is dominated by both ridge estimators, especially when R^2 is small. LOO-CV ridge and minimax ridge have very similar excess prediction risk: LOO-CV is slightly above minimax ridge for the lower two values of R^2 and slightly below minimax ridge when $R^2 = 50\%$.

We next evaluate our approximation to the excess prediction risk of ridge estimators presented in Theorem 2 under DGP-2. Figure 6 displays a comparison between the Monte Carlo estimate of the exact excess prediction risk, $n\left\{R_n(\widehat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2\right\}$ (solid orange line), and our *feasible* approximate excess risk $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ (shaded region). To capture its range of values, we report the central 95 percent range after dropping the lowest 2.5 percent and highest 2.5 percent of values for

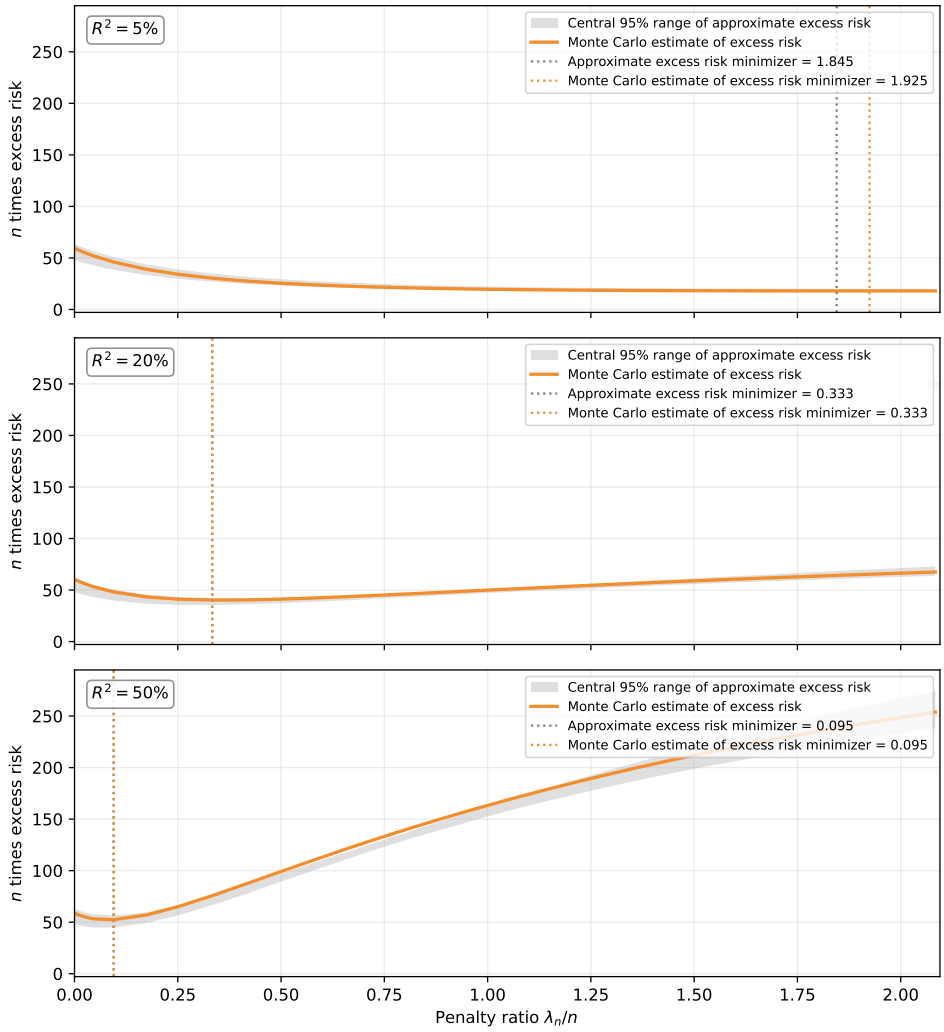

Figure 6: Risk curves under DGP-2

each candidate tuning ratio. In this lower-dimensional version of the design, the feasible approximate excess-risk calculation tracks the Monte Carlo risk curve closely, including the location of the low-risk region. Appendix B.2.2 reports additional diagnostics for the eigenvalues of Σ_2, the coefficient vectors, and the selected tuning ratios.

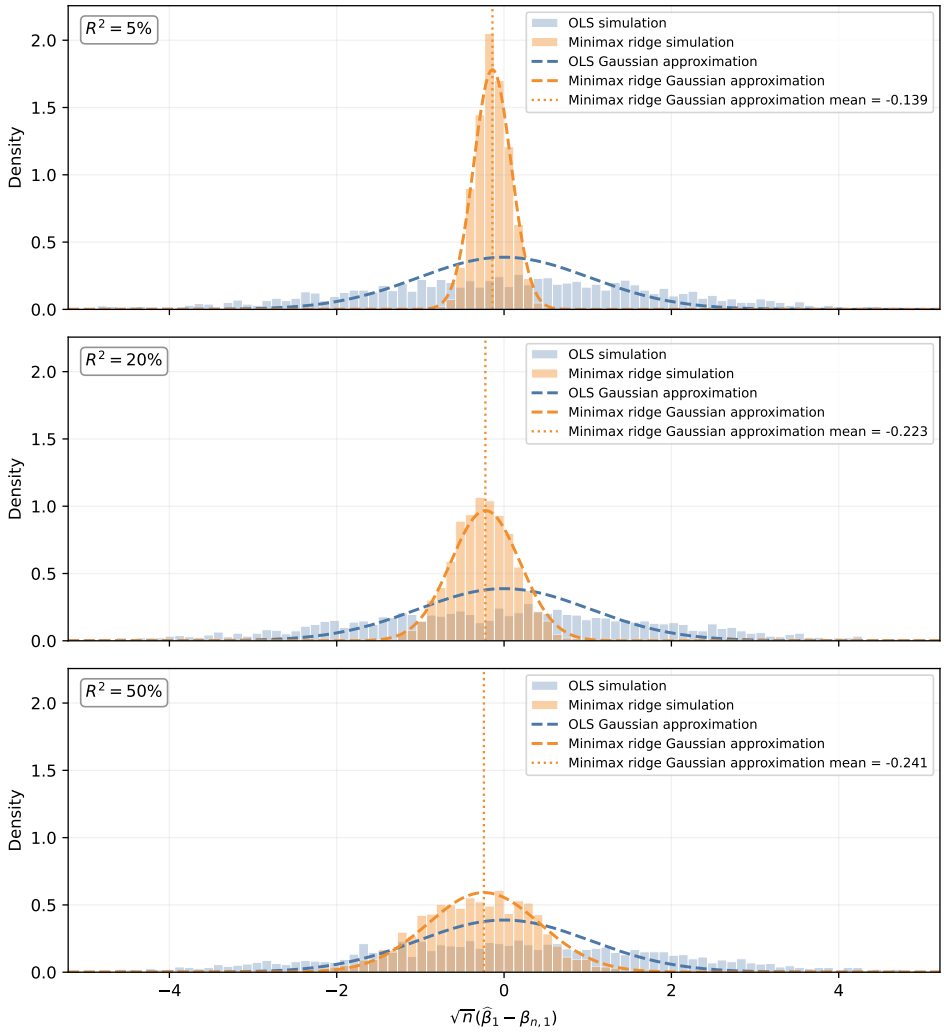

Figure 7: Distribution of the first coefficient under DGP-3

4.3 DGP-3

We next consider the high-dimensional version of the weak-signal design based on Section 3.1 of Shen and Xiu (2025). Our asymptotic theory focused on models in which the number of covariates is small relative to the sample size. Thus, it is of interest to understand the performance of our approximations in models where the covariates are roughly of the same order as the sample size. The setup is exactly the same as the setup in DGP-2 in the previous subsection, except here we set $k = 300$.

The following figures report our simulation results. Figure 7 reports the simulated distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_{n,1})$ for OLS and minimax ridge, together with the Gaussian approximations implied by our theory. The Gaussian curves are computed using the population matrices in the simulation, rather than fitted to the Monte Carlo histograms. Similar to the pattern observed in DGP-1 and DGP-2, comparing with the OLS results, the distribution of the minimax ridge estimator is biased by shrinkage but has a smaller variance.

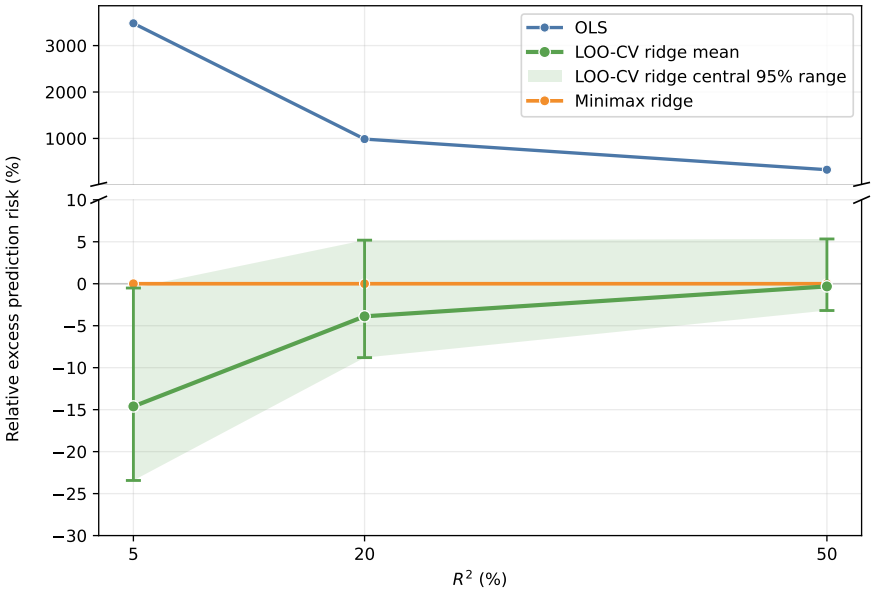

Figure 8: Relative excess prediction risk under DGP-3, LOO-CV and minimax

We next compare the three estimators using excess prediction risk. Figure 8 reports the excess prediction risk of OLS and LOO-CV ridge relative to minimax ridge. The shaded region reports the central 95 percent range of the repetition-specific relative excess risk of LOO-CV ridge. The figure shows that OLS is dominated by both ridge estimators, especially when R^2 is small. LOO-CV ridge and minimax ridge have comparable excess prediction risk, although LOO-CV ridge has slightly lower mean excess risk than minimax ridge in this high-dimensional design. Thus, in the regime where k/n is large, the approximation-based rule should not be interpreted as uniformly improving on LOO-CV, but it remains competitive with the standard cross-validation benchmark.

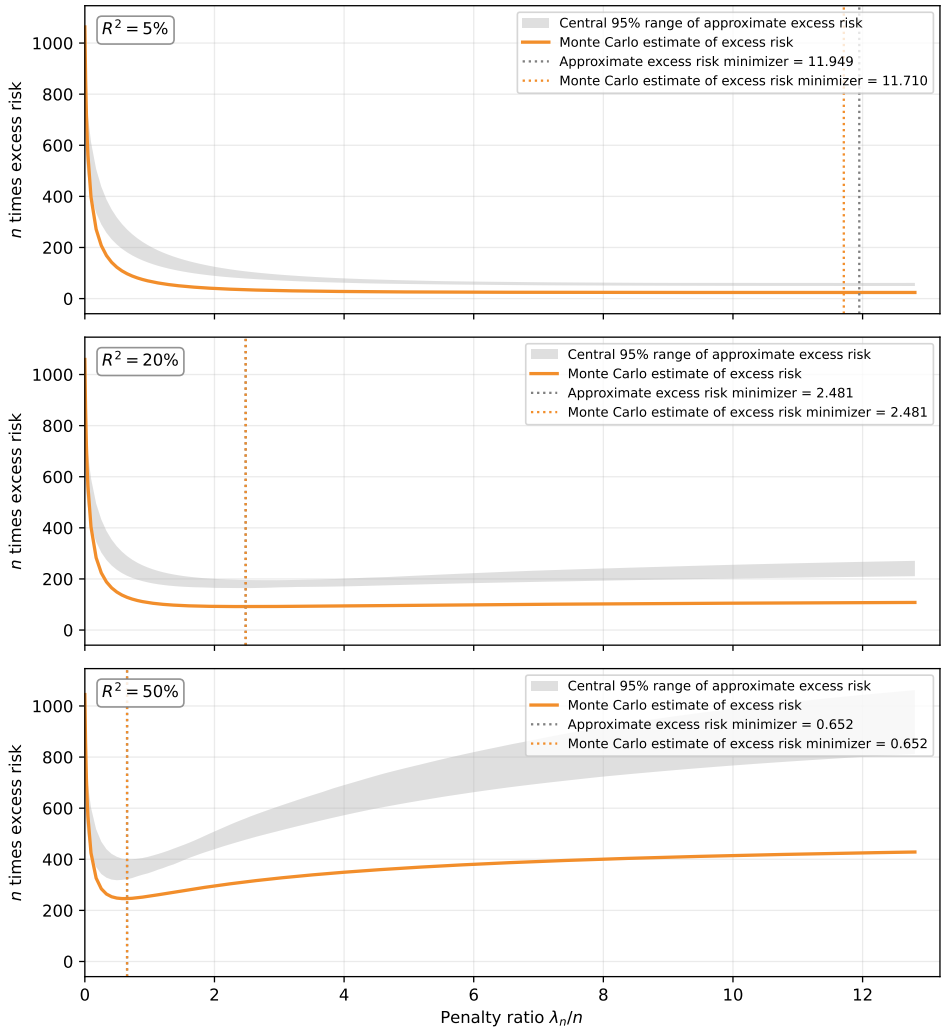

Figure 9: Risk curves under DGP-3

We next evaluate our approximation to the excess prediction risk of ridge estimators presented in Theorem 2 under DGP-3. Figure 9 displays a comparison between the Monte Carlo estimate of the exact excess prediction risk, $n\{R_n(\widehat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2\}$ (solid orange line), and our *feasible* approximate excess risk $R_n^e(\lambda; b, \widehat{\Sigma}, \widehat{\Omega})$ (shaded region). To capture its range of values, we report the central 95 percent range after dropping the lowest 2.5 percent and highest 2.5 percent of values for each candidate tuning ratio. In this high-dimensional version of the design, the feasible approximation is less accurate than in DGP-2, especially away from the low-risk region. This pattern is

expected because $k/n = 0.6$ is far from the low-dimensional asymptotic framework used to derive our approximation. Appendix B.2.3 reports additional diagnostics for the eigenvalues of Σ_2, the coefficient vectors, and the selected tuning ratios.

5 Conclusion

We presented a simple Gaussian approximation to the finite-sample distribution of the ridge regression estimator. Our approximation is based on nonstandard asymptotics where *i*) we let the estimator's regularization parameter grow proportionally to the sample size; and *ii*) we treat the population regression coefficients as *local* to the reference vector that defines the estimator's direction of shrinkage. In contrast to other asymptotic approximations available in the literature, we allow for general forms of heteroskedasticity and autocorrelation in the data generating process (at the cost of considering a low-dimensional model where covariates are not allowed to grow with the sample size). We used our simple Gaussian approximation to propose two new strategies to select the regularization parameter for the ridge regression estimator. The suggested strategies select the regularization parameter to minimize either average or worst-case excess prediction risk, where risk is computed using our suggested Gaussian approximation.

References

Abadie, A. and M. Kasy (2019): "Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning," *The Review of Economics and Statistics*, 101, 743–762.

Andrews, I. and A. Mikusheva (2022): "Optimal decision rules for weak GMM," *Econometrica*, 90, 715–748.

Anup, A. K. and G. Maddala (1984): "Ridge estimators for distributed lag models," *Communications in Statistics-Theory and Methods*, 13, 217–225.

Atanasov, A., J. A. Zavatone-Veth, and C. Pehlevan (2024): "Risk and cross validation in ridge regression with correlated samples," *arXiv preprint arXiv:2408.04607*.

Brockwell, P. J. and R. A. Davis (2013): *Time series: theory and methods*, Springer Science & Business Media.

Cambanis, S., S. Huang, and G. Simons (1981): "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, 11, 368–385.

DasGupta, A. (2008): *Asymptotic Theory of Statistics and Probability*, Springer Verlag.

Dobriban, E. and S. Wager (2018): "High-dimensional asymptotics of prediction: Ridge regression and classification," *The Annals of Statistics*, 46, 247–279.

Dou, L. and U. K. Müller (2021): "Generalized Local-to-Unity Models," *Econometrica*, 89, 1825–1854.

Ferguson, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

Friedman, J., T. Hastie, and R. Tibshirani (2017): *The elements of statistical learning: data mining, inference and prediction*, vol. 1 of *Series in Statistics*, New York: Springer, second edition ed.

Gibbons, D. G. (1981): "A Simulation Study of Some Ridge Estimators," *Journal of the American Statistical Association*, 76, 131–139.

Hansen, B. (2022): *Econometrics*, Princeton University Press.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022): "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, 50, 949–986.

Hoerl, A. E. (1962): "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, 58, 54–59.

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

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

Hsu, D., S. M. Kakade, and T. Zhang (2012): "Random design analysis of ridge regression," in *Conference on learning theory*, JMLR Workshop and Conference Proceedings, 9–1.

Knight, K. and W. Fu (2000): "Asymptotics for lasso-type estimators," *Annals of statistics*, 1356–1378.

Montiel Olea, J. L., C. Rush, A. Velez, and J. Wiesel (2026): "The distributionally robust prediction error of the LASSO and related estimators," *The Annals of Statistics*, 54, 1006–1027.

Mourtada, J. (2022): "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices," *The Annals of Statistics*, 50, 2157–2178.

Mourtada, J. and L. Rosasco (2022): "An elementary analysis of ridge regression with random design," *Comptes Rendus. Mathématique*, 360, 1055–1063.

Patil, P., J.-H. Du, and R. J. Tibshirani (2024): "Optimal ridge regularization for out-of-distribution prediction," *arXiv preprint arXiv:2404.01233*.

Patil, P., Y. Wei, A. Rinaldo, and R. Tibshirani (2021): "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in *International conference on artificial intelligence and statistics*, PMLR, 3178–3186.

Phillips, P. C. (1988): "Regression theory for near-integrated time series," *Econometrica: Journal of the Econometric Society*, 1021–1043.

Powell, J. L. (2017): "Identification and Asymptotic Approximations: Three Examples of Progress in Econometric Theory," *Journal of Economic Perspectives*, 31, 107–124.

Shen, Z. and D. Xiu (2025): "Can Machines Learn Weak Signals?" Working Paper 33421, National Bureau of Economic Research, Cambridge, MA.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Swindel, B. F. (1976): "Good ridge estimators based on prior information," *Communications in Statistics-Theory and Methods*, 5, 1065–1075.

Velez, A. (2024): "On the asymptotic properties of debiased machine learning estimators," *arXiv preprint arXiv:2411.01864*.

A Proofs of Main Results

A.1 Proof of Proposition 1

Proof. Let X be the $n \times k$ matrix that contains $x_i^\top$ in its i-th row. Let Y and ϵ be the $n \times 1$ matrices that contain y_i and ϵ_i (respectively) in their i-th row. Define the matrices $\widehat{\Sigma} \equiv X^\top X/n$ and $\widehat{A} \equiv \widehat{\Sigma} + (\lambda_n/n)\mathbb{I}_k$. Using this notation, $\widehat{\beta}_{\lambda_n} = \widehat{A}^{-1}\left(X^\top Y/n + (\lambda_n/n)\beta_0\right)$ and $\widehat{\beta}_{OLS} = \widehat{\Sigma}^{-1}(X^\top Y)/n$. Algebra shows

$$\begin{aligned} \sqrt{n}\left(\widehat{\beta}_{\lambda_n} - \beta_n\right) &= \widehat{A}^{-1}\left(X^\top \epsilon/\sqrt{n} - (\lambda_n/n)\sqrt{n}\,(\beta_n - \beta_0)\right) \\ \sqrt{n}\left(\widehat{\beta}_{OLS} - \beta_n\right) &= \widehat{\Sigma}^{-1} X^\top \epsilon/\sqrt{n}\ . \end{aligned} \tag{30}$$

Therefore,

$$\sqrt{n}\left(\widehat{\beta}_{\lambda_n} - \beta_n\right) - \sqrt{n}\left(\widehat{\beta}_{OLS} - \beta_n\right) = \left(\widehat{A}^{-1} - \widehat{\Sigma}^{-1}\right)(X^\top \epsilon)/\sqrt{n} - \widehat{A}^{-1}(\lambda_n/n)\sqrt{n}\,(\beta_n - \beta_0)\ ,$$

which is $o_p(1)$ because (i) $\lambda_n/n \xrightarrow{p} 0$ implies $\widehat{A}^{-1} - \widehat{\Sigma}^{-1}$ is $o_p(1)$, (ii) Assumption 1 implies $(X^\top \epsilon)/\sqrt{n}$ is $O_p(1)$, and (iii) because we are assuming $(\lambda_n/n)\sqrt{n}\,(\beta_n - \beta_0) \to 0$ in this proposition. □

A.2 Proof of Theorem 1

Proof. Recall that $\widehat{\Sigma} \equiv X^\top X/n$ and $\widehat{A} \equiv \widehat{\Sigma} + (\lambda_n/n)\mathbb{I}_k$. Note that $\widehat{A} \xrightarrow{p} \Sigma + \lambda\mathbb{I}_k$ since $\widehat{\Sigma} \xrightarrow{p} \Sigma$ (Assumption 1) and $\lambda_n/n \xrightarrow{p} \lambda$. Note also that $\widehat{A}^{-1}(\lambda_n/n)\sqrt{n}\,(\beta_n - \beta_0) \xrightarrow{p} (\Sigma + \lambda\mathbb{I}_k)^{-1}\lambda b$. We conclude by using Equation (30), Assumption 1, and Slutsky's theorem. □

A.3 Proof of Theorem 2

Proof. We establish the result in two steps. We first show that the excess prediction risk $R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n)$ is approximated by $R_n^e(\lambda, b, \Sigma, \Omega)$ up to a remainder error of size $o(n^{-1})$. We then show that

$R_n^e(\lambda, b, \Sigma, \Omega)$, which depends on population parameters (Σ, Ω), can be approximated by $R_n^e(\lambda, b, \widehat{\Sigma}, \widehat{\Omega})$, which uses consistent estimators $(\widehat{\Sigma}, \widehat{\Omega})$, up to error of size $o_{(\beta_n, \mathbb{P}_n)}(1/n)$. The conclusion follows by combining these two steps and the triangular inequality.

Step 1: For a new draw $(x^\top, \epsilon)$ from the stationary distribution $\mathbb{P}$ associated to $\mathbb{P}_n$, we use β_n to construct a new outcome-covariate pair $(y, x) = (x^\top \beta_n + \epsilon, x)$. The exact finite-sample prediction risk of $\hat{a}_{\lambda_n}$ is:

$$R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) = \mathbb{E}_{(\beta_n, \mathbb{P}_n)}\big[(y - x^\top \widehat{\beta}_{\lambda_n})^2\big] = \sigma^2 + \mathbb{E}_{(\beta_n, \mathbb{P}_n)}\big[(\widehat{\beta}_{\lambda_n} - \beta_n)^\top \Sigma (\widehat{\beta}_{\lambda_n} - \beta_n)\big].$$

Defining $Z_n \equiv \sqrt{n}(\widehat{\beta}_{\lambda_n} - \beta_n)$, this becomes

$$R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) = \sigma^2 + \frac{1}{n}\mathbb{E}_{(\beta_n, \mathbb{P}_n)}\big[Z_n^\top \Sigma Z_n\big].$$

Note that equation 18 can be rewritten as

$$R_n^e(\lambda, b, \Sigma, \Omega) = \frac{1}{n}\mathbb{E}_Z[Z^\top \Sigma Z],$$

where

$$Z \sim \mathcal{N}_k(-(\Sigma + \lambda \mathbb{I}_k)^{-1}\lambda b, (\Sigma + \lambda \mathbb{I}_k)^{-1}\Omega(\Sigma + \lambda \mathbb{I}_k)^{-1}).$$

Therefore,

$$R_n(\hat{a}_{\lambda_n}; \beta_n, \mathbb{P}_n) - \sigma^2 - R_n^e(\lambda, b, \Sigma, \Omega) = \frac{1}{n}\Big(\mathbb{E}_{(\beta_n, \mathbb{P}_n)}\big[Z_n^\top \Sigma Z_n\big] - \mathbb{E}_Z[Z^\top \Sigma Z]\Big),$$

so it suffices to show

$$\mathbb{E}_{(\beta_n, \mathbb{P}_n)}\big[Z_n^\top \Sigma Z_n\big] \to \mathbb{E}_Z[Z^\top \Sigma Z].$$

Under the assumptions of Theorem 2, Slutsky's theorem and Theorem 1 imply that $Z_n \xrightarrow{d} Z$.

Theorem 6.2 in DasGupta (2008) implies that under assumption iii) of Theorem 2 (which is a sufficient condition for uniform integrability) we have $\mathbb{E}_{(\beta_n,\mathbb{P}_n)}\big[Z_n^\top \Sigma Z_n\big] \to \mathbb{E}_Z[Z^\top \Sigma Z]$. This establishes the oracle approximation

$$R_n(\hat{a}_{\lambda_n};\beta_n,\mathbb{P}_n) - \sigma^2 = R_n^e(\lambda,b,\Sigma,\Omega) + o\Big(\frac{1}{n}\Big).$$

Step 2: It is sufficient to show that $n\left(R_n^e(\lambda,b,\widehat{\Sigma},\widehat{\Omega}) - R_n^e(\lambda,b,\Sigma,\Omega)\right) = o_{(\beta_n,\mathbb{P}_n)}(1)$. We claim that this follows from the consistency of the estimators $(\widehat{\Sigma},\,\widehat{\Omega})$ of $(\Sigma,\,\Omega)$. To see this, note that

$$nR_n^e(\lambda;b,\Sigma,\Omega) \equiv \lambda^2 b^\top(\Sigma+\lambda\mathbb{I}_k)^{-1}\Sigma(\Sigma+\lambda\mathbb{I}_k)^{-1}b + \operatorname{trace}\left((\Sigma+\lambda\mathbb{I}_k)^{-1}\Omega(\Sigma+\lambda\mathbb{I}_k)^{-1}\Sigma\right).$$

Then, we conclude by using Slutsky's theorem and continuous mapping theorem. □

A.4 Proof of Theorem 3

Proof. Under isotropic features, $\Sigma = \sigma_x^2 I_k$, the approximation of the excess risk of ridge regression in Equation (18) simplifies to

$$R_n^e(\lambda;b,\sigma_x^2,\Omega) = \frac{1}{n}\frac{\lambda^2\sigma_x^2}{(\sigma_x^2+\lambda)^2}b^\top b + \frac{1}{n}\frac{\sigma_x^2}{(\sigma_x^2+\lambda)^2}\operatorname{trace}(\Omega). \tag{31}$$

We use the previous closed-form expression to obtain

$$\partial_\lambda R_n^e(\lambda;b,\sigma_x^2,\Omega) = \frac{1}{n}\frac{2\sigma_x^2}{(\sigma_x^2+\lambda)^3}(\sigma_x^2(b^\top b)\lambda - \operatorname{trace}(\Omega)). \tag{32}$$

Note that the previous expression is negative if and only if $\lambda < \lambda^* \equiv \frac{\operatorname{trace}(\Omega)}{\sigma_x^2(b^\top b)}$. Therefore, $R_n^e(\lambda;b,\sigma_x^2,\Omega)$ as a function of λ is decreasing on $[0,\lambda^*]$ and increasing on $[\lambda^*,+\infty]$. As a result, we have that $\lambda^* \in \arg\min_{\lambda\geq 0} R_n^e(\lambda;b,\sigma_x^2,\Omega)$ for any given (b,σ_x^2,Ω).

We first establish part (1) of Theorem 3. Note that when $\|\cdot\|$ is Euclidean norm, the set

$\mathcal{B} \equiv \{b \in \mathbb{R}^k \mid \|b\| \leq B\} = \{b \in \mathbb{R}^k \mid b^\top b \leq B^2\}$. Since the coefficient on $b^\top b$ in the excess risk approximation formula defined by Equation (31) is nonnegative for every $\lambda \geq 0$, the worst-case value of the approximate excess risk is attained when $b^\top b = B^2$. Hence the minimax problem reduces to minimizing the function

$$Q_{\text{minimax}}(\lambda) \equiv \frac{1}{n} \frac{\lambda^2 \sigma_x^2}{(\sigma_x^2 + \lambda)^2} B^2 + \frac{1}{n} \frac{\sigma_x^2}{(\sigma_x^2 + \lambda)^2} \operatorname{trace}(\Omega), \qquad \lambda \geq 0,$$

which as a function of λ achieves its minimum at $\lambda^*_{\text{minimax}} = \frac{\operatorname{trace}(\Omega)}{\sigma_x^2 B^2}$.

We then establish part (2) of Theorem 3. Suppose $\mathbb{E}_\pi[b^\top b] < \infty$. Taking expectation with respect to π gives

$$Q_{\text{average}}(\lambda) := \mathbb{E}_\pi\big[R_n(\lambda; b, \sigma_x^2, \Omega)\big] = \frac{1}{n} \frac{\lambda^2 \sigma_x^2}{(\sigma_x^2 + \lambda)^2} \mathbb{E}_\pi[b^\top b] + \frac{1}{n} \frac{\sigma_x^2}{(\sigma_x^2 + \lambda)^2} \operatorname{trace}(\Omega), \qquad \lambda \geq 0,$$

which as a function of λ achieves its minimum at $\lambda^*_{\text{average}} = \frac{\operatorname{trace}(\Omega)}{\sigma_x^2 \mathbb{E}_\pi[b^\top b]}$. □

B Additional Results

B.1 Optimal choice of λ for elliptical contoured distributions

In this appendix, we extend the result on choosing λ to minimize average approximate excess risk to the more general case in which b follows an elliptical contoured distribution. Following Cambanis, Huang, and Simons (1981), a k-dimensional random vector X is said to follow an elliptical distribution with location parameter $\mu \in \mathbb{R}^k$, nonnegative definite matrix $\Sigma \in \mathbb{R}^{k \times k}$, and characteristic generator ψ if the characteristic function of $X - \mu$ satisfies

$$\phi_{X-\mu}(t) = \psi(t^\top \Sigma t), \qquad t \in \mathbb{R}^k.$$

We denote this by $X \sim EC_k(\mu, \Sigma, \psi)$.

Now suppose $b \sim EC_k(0, I_k, \psi)$ has finite second moments. Algebra shows that

$$\mathbb{E}[bb^\top] = \frac{\mathbb{E}\|b\|^2}{k} I_k.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{b\sim\pi}\Big[b^\top(\widehat{\Sigma}+\lambda I_k)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_k)^{-1}b\Big] \\
&\quad= \mathbb{E}_{b\sim\pi}\Big[\mathrm{tr}\Big((\widehat{\Sigma}+\lambda I_k)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_k)^{-1}bb^\top\Big)\Big] \\
&\quad= \mathrm{tr}\Big((\widehat{\Sigma}+\lambda I_k)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_k)^{-1}\mathbb{E}_{b\sim\pi}[bb^\top]\Big) \\
&\quad= \mathrm{tr}\left((\widehat{\Sigma}+\lambda I_k)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_k)^{-1}\frac{\mathbb{E}\|b\|^2}{k} I_k\right) \\
&\quad= \frac{\mathbb{E}\|b\|^2}{k}\,\mathrm{tr}\Big((\widehat{\Sigma}+\lambda I_k)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_k)^{-1}\Big).
\end{aligned}$$

Hence the average approximate excess risk takes the same form as in Section 3.3.2, with C^2 replaced by $\mathbb{E}\|b\|^2/k$. It follows that the corresponding characterization of the optimal choice of λ extends immediately to the class of elliptical contoured distributions with finite second moment.

B.2 Additional Simulation Results

B.2.1 DGP-1

DGP-1 is the isotropic example. The main additional object of interest is the selected tuning ratio $\widehat{\lambda}_n/n$. Under the population version of the minimax rule in this design, the optimal scaled regularization parameter equals

$$c^* = \frac{\mathrm{trace}(\Omega)}{\sigma_x^2 B^2} = 1.$$

Figure 10 reports the distribution of the selected tuning ratio $\widehat{\lambda}_n/n$ for LOO-CV ridge and Minimax ridge. The feasible minimax rule is tightly centered near the population value $c^* = 1$, with dispersion decreasing as the sample size increases. In contrast, the LOO-CV tuning ratio is

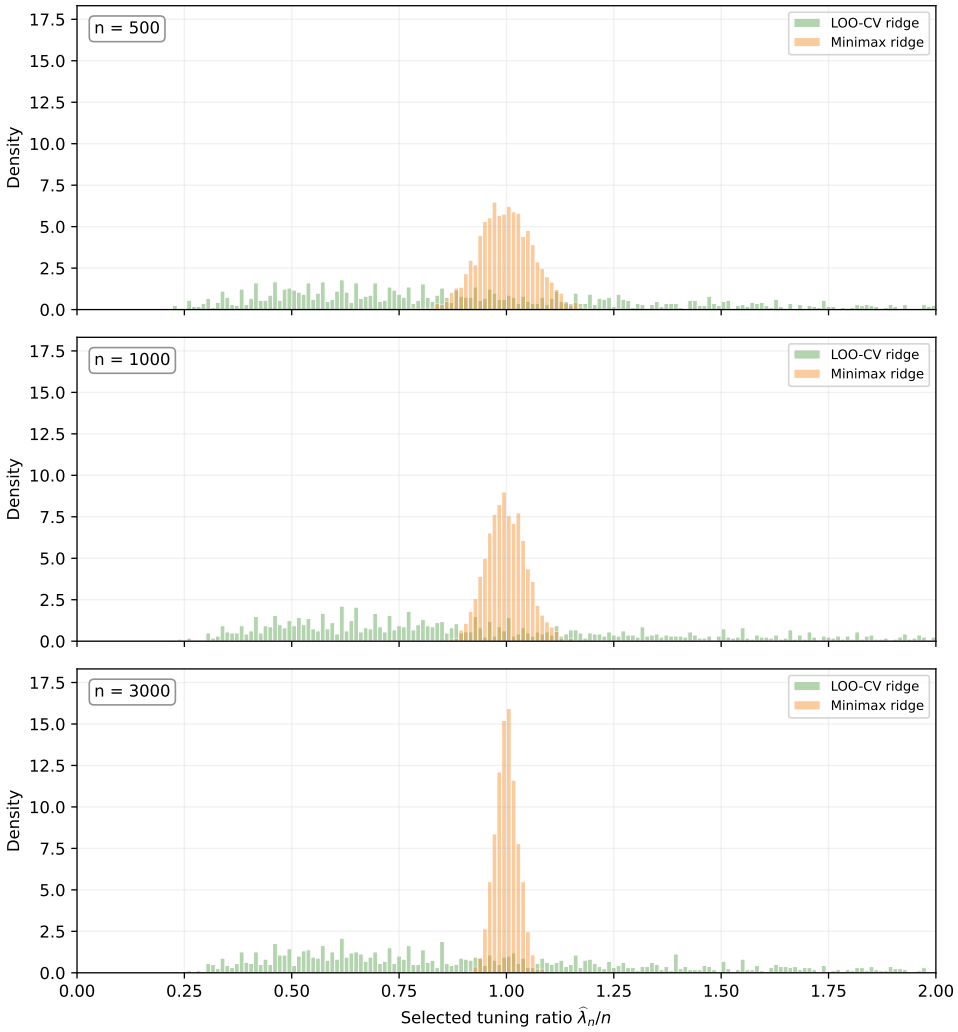

Figure 10: Scaled regularization parameter $(\widehat{\lambda}_n/n)$ used by the LOO-CV and minimax ridge estimators under DGP-1

much more dispersed across Monte Carlo repetitions. This helps explain the risk comparison: in DGP-1, minimax ridge uses a tuning parameter close to the risk-minimizing value, whereas LOO-CV often selects substantially different amounts of shrinkage.

B.2.2 DGP-2

This appendix records diagnostic information for the covariance matrix and coefficient vectors used in DGP-2. These diagnostics are useful because the design departs from the isotropic benchmark:

the prediction covariance is Σ_2, not a scalar multiple of the identity matrix, and the coefficient vector is a fixed draw from the spike-and-slab model described in the main text.

First, we verify the equality $\Sigma = \mathbb{E}[x_i x_i^\top] = \Sigma_2$ used in the simulation section. Let e_i denote the ith canonical basis vector in $\mathbb{R}^n$. Since the ith row of X is $x_i^\top = e_i^\top X$, we have

$$\begin{aligned}\mathbb{E}[x_i x_i^\top] &= \mathbb{E}[X^\top e_i e_i^\top X] \\ &= \Sigma_2^{1/2}\,\mathbb{E}\Big[Z^\top \Sigma_1^{1/2} e_i e_i^\top \Sigma_1^{1/2} Z\Big]\,\Sigma_2^{1/2}.\end{aligned}$$

For any deterministic $n \times n$ matrix A and an $n \times k$ matrix Z with i.i.d. standard normal entries, $\mathbb{E}[Z^\top A Z] = \operatorname{trace}(A) I_k$. Applying this identity with $A = \Sigma_1^{1/2} e_i e_i^\top \Sigma_1^{1/2}$ gives

$$\begin{aligned}\mathbb{E}[x_i x_i^\top] &= \Sigma_2^{1/2} \operatorname{trace}\Big(\Sigma_1^{1/2} e_i e_i^\top \Sigma_1^{1/2}\Big)\, I_k \Sigma_2^{1/2} \\ &= \Sigma_2^{1/2} (e_i^\top \Sigma_1 e_i) I_k \Sigma_2^{1/2} \\ &= \Sigma_2^{1/2} (\Sigma_1)_{ii} I_k \Sigma_2^{1/2} \\ &= \Sigma_2,\end{aligned}$$

where the last equality uses $(\Sigma_1)_{ii} = 1$. Similarly, because the errors are independent of the regressors, mean zero, and have fixed variances $\bar{\sigma}^2_{\epsilon,i}$, the corresponding variance matrix for the score is proportional to Σ_2:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i x_i^\top \epsilon_i^2] = \left(\frac{1}{n}\sum_{i=1}^n \bar{\sigma}^2_{\epsilon,i}\right)\Sigma_2.$$

Figure 11 plots the eigenvalues of the fixed Σ_2 matrix used in DGP-2. The eigenvalues range from 0.502 to 1.479, with mean 0.981 and max-min spread 0.977. The vertical dashed line marks the isotropic benchmark value one. The figure shows that the design is centered near the isotropic benchmark on average but has meaningful dispersion in eigenvalues.

Figure 12 displays the fixed coefficient vector used in DGP-2 after rescaling the same preliminary draw to the three target values of R^2. The labels report $\|\beta_n\|$ for each target value. In this draw,

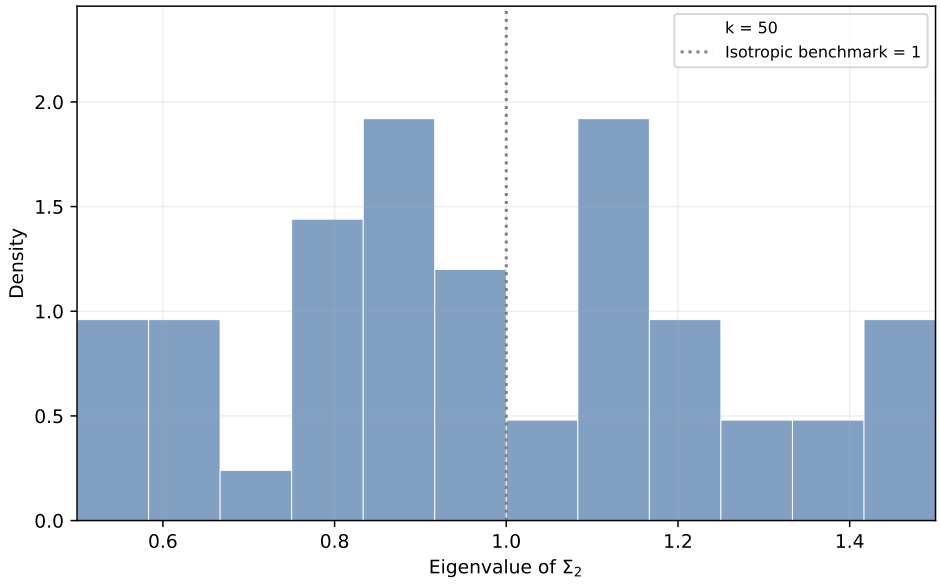

Figure 11: Eigenvalues of Σ_2 under DGP-2

41 out of the 50 coordinates are nonzero. The figure makes explicit how increasing the target R^2 changes the magnitude of the coefficient vector while preserving the same coefficient pattern.

Figure 13 reports the distribution of the selected tuning ratio $\widehat{\lambda}_n/n$ for LOO-CV ridge and minimax ridge under DGP-2. The minimax rule is more concentrated because it is driven by the approximate-risk calculation and the calibrated radius B, whereas LOO-CV is more dispersed across Monte Carlo repetitions.

B.2.3 DGP-3

This appendix records diagnostic information for the covariance matrix and coefficient vectors used in DGP-3. DGP-3 uses the same covariance-generation and coefficient-generation schemes as DGP-2, but increases the number of covariates from $k = 50$ to $k = 300$ while keeping $n = 500$. Thus, the diagnostics below should be read as the high-dimensional counterpart to Appendix B.2.2.

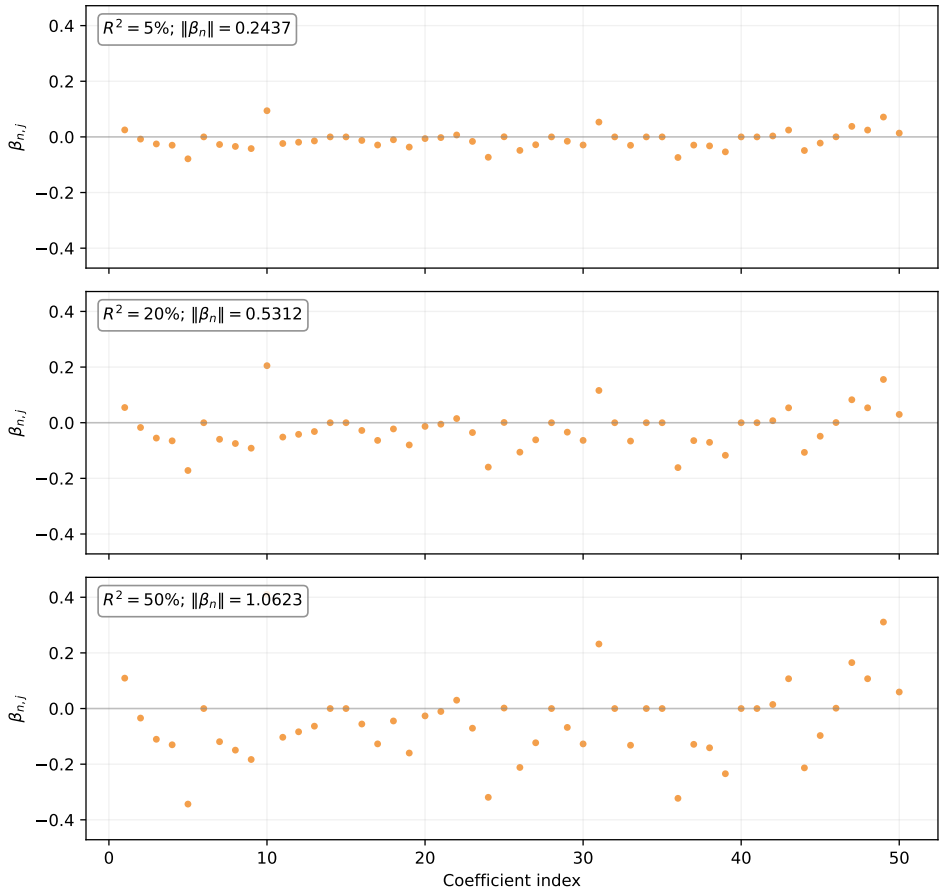

Figure 12: Coefficient vectors under DGP-2

The same algebra as in Appendix B.2.2 implies that the population prediction covariance is

$$\Sigma = \mathbb{E}[x_i x_i^\top] = \Sigma_2,$$

because the diagonal entries of Σ_1 are equal to one. Since the errors are independent of the regressors, mean zero, and have fixed variances $\bar{\sigma}^2_{\epsilon,i}$, the corresponding score variance matrix is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[x_i x_i^\top \epsilon_i^2] = \left(\frac{1}{n}\sum_{i=1}^{n} \bar{\sigma}^2_{\epsilon,i}\right) \Sigma_2.$$

Figure 14 plots the eigenvalues of the fixed Σ_2 matrix used in DGP-3. The eigenvalues range

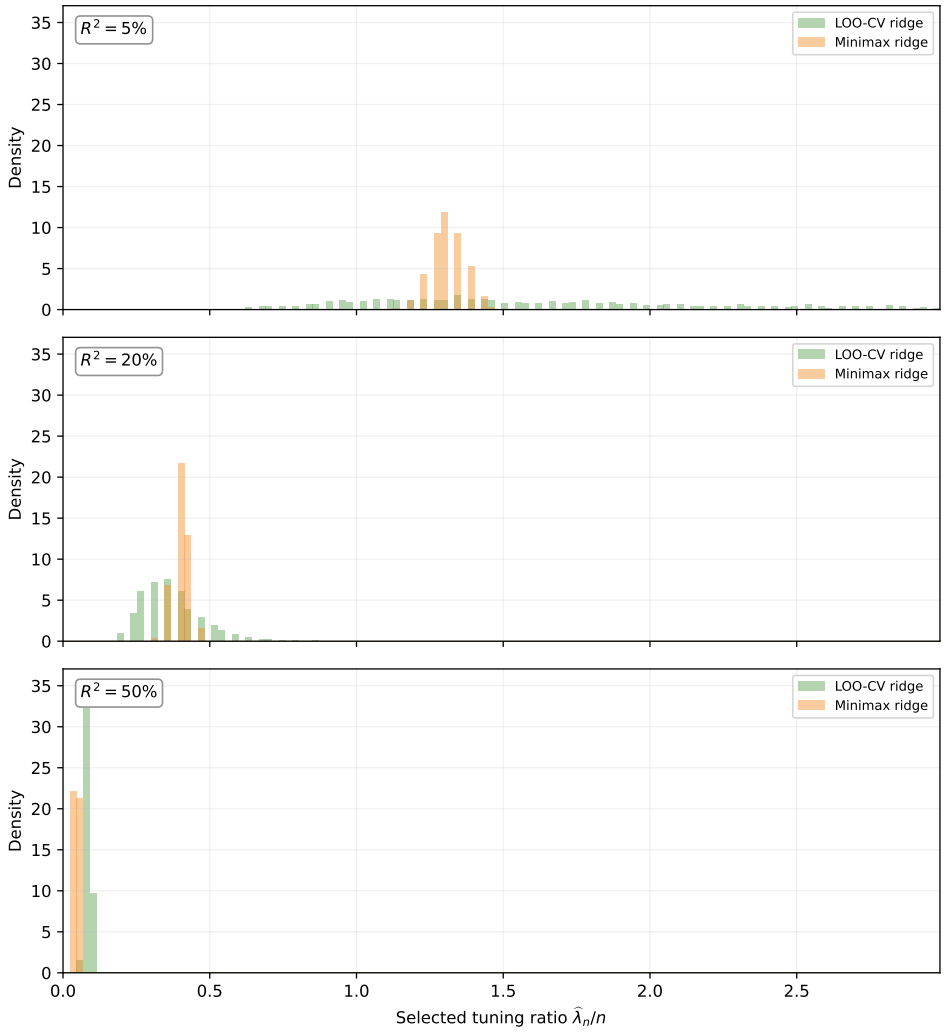

Figure 13: Selected tuning ratios under DGP-2

from 0.502 to 1.499, with mean 1.027 and max-min spread 0.997. As in DGP-2, the design is centered near the isotropic benchmark on average but is clearly non-isotropic.

Figure 15 displays the fixed coefficient vector used in DGP-3 after rescaling the same preliminary draw to the three target values of R^2. The labels report $\|\beta_n\|$ for each target value. In this draw, 244 out of the 300 coordinates are nonzero. The figure shows that the signal is dense but individually weak, especially for the lower target value of R^2.

Figure 16 reports the distribution of the selected tuning ratio $\widehat{\lambda}_n/n$ for LOO-CV ridge and

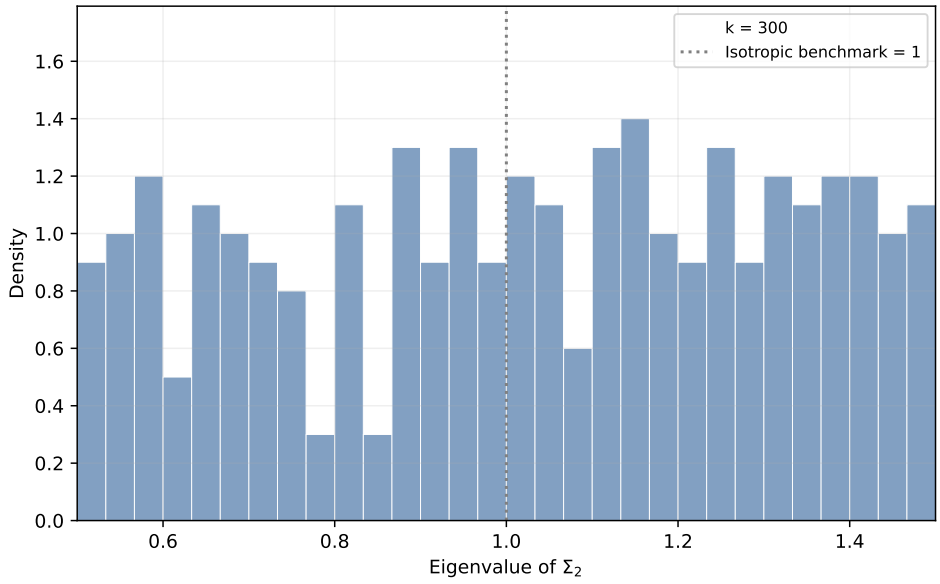

Figure 14: Eigenvalues of Σ_2 under DGP-3

minimax ridge under DGP-3. Relative to DGP-2, the LOO-CV choices are substantially more dispersed, especially when R^2 is small. This dispersion is consistent with the main-text finding that, in the high-dimensional design, LOO-CV and minimax ridge have comparable average risk but can choose very different penalty values in individual samples.

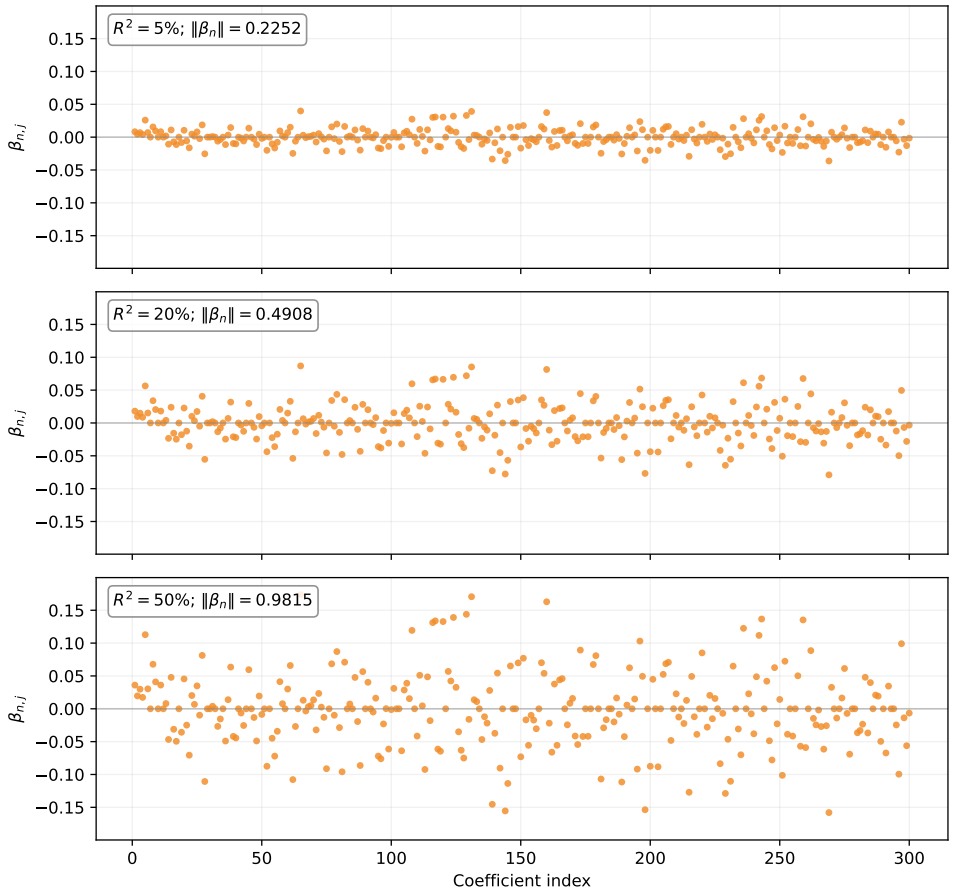

Figure 15: Coefficient vectors under DGP-3

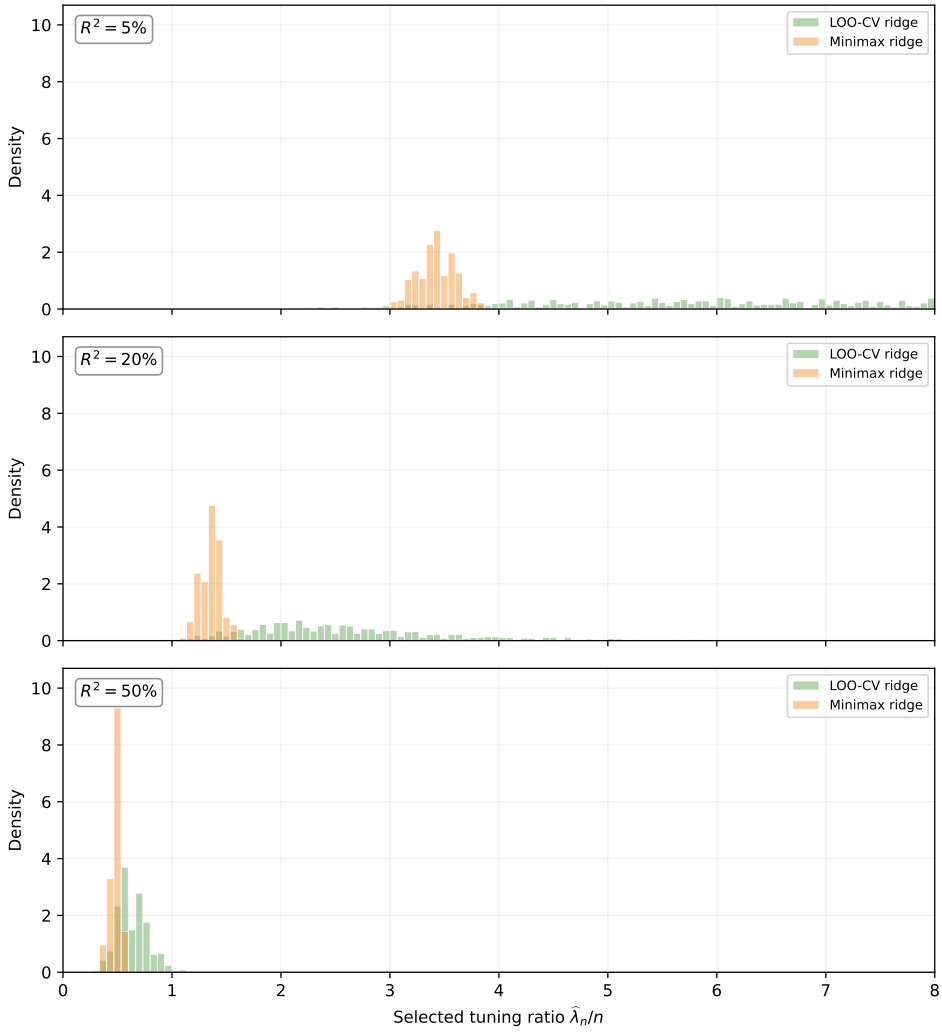

Figure 16: Selected tuning ratios under DGP-3