ROTI-GCV: Generalized Cross-Validation for Right-Rotationally Invariant Data

Kevin Luo
Harvard University

Yufan Li
Harvard University

Pragya Sur
Harvard University

Abstract

Two key tasks in high-dimensional regularized regression are tuning the regularization strength for accurate predictions and estimating the out-of-sample risk. It is known that the standard approach — k-fold cross-validation — is inconsistent in modern high-dimensional settings. While leave-one-out and generalized cross-validation remain consistent in some high-dimensional cases, they become inconsistent when samples are dependent or contain heavy-tailed covariates. As a first step towards modeling structured sample dependence and heavy tails, we use right-rotationally invariant covariate distributions — a crucial concept from compressed sensing. In the proportional asymptotics regime where the number of features and samples grow comparably, which is known to better reflect the empirical behavior in moderately sized datasets, we introduce a new framework, ROTI-GCV, for reliably performing cross-validation under these challenging conditions. Along the way, we propose new estimators for the signal-to-noise ratio and noise variance. We conduct experiments that demonstrate the accuracy of our approach in a variety of synthetic and semi-synthetic settings.

1 INTRODUCTION

Regularized estimators are fundamental for modern high-dimensional statistics. In this context, the prototypical problem of ridge regression has gained renewed attention, with multiple works characterizing its out-of-sample risk in high dimensions, c.f., Dobriban and Wager (2018); Hastie et al. (2022). Additionally, consistent estimates for this out-of-sample risk, crucial for optimizing the regularization parameter λ, have been established in the form of leave-one-out cross-validation (LOOCV) and generalized cross-validation (GCV) (Xu et al., 2021; Patil et al., 2021; Atanasov et al., 2024b).

Crucially, all aforementioned analyses assume that samples are independent and identically distributed (i.i.d.) draws from some distribution with sufficient regularity, such as light tails. These assumptions often fail in practice, such as in financial returns, neuroscience, and climate data (Grobys, 2021; Zscheischler et al., 2021; Tagliazucchi et al., 2013). In such cases, cross-validation techniques can become biased (Theorem 1, Remark 3), necessitating alternatives. Only recently, Bigot et al. (2024) considered independent but non-identically distributed samples and Atanasov et al. (2024a) explored Gaussian designs with both sample and feature dependence. Understanding cross-validation in high dimensions under more general sample dependencies and for designs with heavier tails than sub-Gaussians remains a significant challenge. This paper takes the first step in addressing this issue.

We assess the impact of sample dependence and heavy tails in ridge regression and cross-validation using an alternative random matrix ensemble for the covariate distribution. Instead of assuming a Gaussian matrix or i.i.d. rows from a fixed distribution, we require the singular value decomposition of the design $\mathbf{X}$ to be *right-rotationally invariant* (see Definition 1). We characterize the ridge regression risk for these designs and introduce ROTI-GCV, a new GCV-inspired framework for consistently estimating the out-of-sample risk. Simulations show that ROTI-GCV outperforms traditional GCV and LOOCV in situations with correlated and heavy-tailed observations.

Right-rotationally invariant designs facilitate tractable analysis while capturing structured forms of sample dependencies and heavy-tailed covariates. These designs have gained significant attention as alternatives to i.i.d. designs for theoretical analysis in numerous studies (Takeda et al., 2006; Ma and Ping, 2017; Rangan et al.,

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

2019; Fan, 2022; Gerbelot et al., 2020; Takeuchi, 2019; Liu et al., 2022; Li et al., 2023; Fan and Wu, 2021; Xu et al., 2022). Recent works (Dudeja et al., 2023; Wang et al., 2023) showed that, under mild conditions, properties of convex estimators across various designs are well-approximated by those under a right-rotationally invariant design with the same spectrum, indicating a broad universality class for these distributions.

Formally, we operate in the proportional asymptotics regime, where the dimension p scales proportionally with the sample size n. This regime, rooted in probability theory and statistical physics, has received significant attention due to its ability to accurately capture empirical phenomena (Hastie et al., 2022; Sur and Candès, 2019; Sur et al., 2019; Liang and Sur, 2022). Furthermore, results derived under this regime require minimal assumptions on the signal structure, resulting in theory and methods with broad practical utility Song et al. (2024). In regards to cross-validation for i.i.d. samples, Hastie et al. (2022); Rahnama Rad and Maleki (2018); Wang et al. (2018) established that classical k-fold cross-validation inconsistently estimates the out-of-sample error in this high-dimensional regime, whereas LOOCV and GCV remain consistent, prompting our study of GCV. However, our work focuses on non-i.i.d. samples and non-Gaussian tailed designs, areas not sufficiently addressed in prior work.

We provide proofs for ridge regression, but emphasize that our core idea can extend to other penalties. Our proof hinges on characterizing the limit of GCV (Theorem 1). This reveals that the original GCV is inconsistent for the out-of-sample risk in our setting. We next observe that though GCV is inconsistent, its asymptotic limit can serve as an estimating equation for the unknown parameters parameterizing the risk. This estimating equation approach directly inspires a new cross-validation method that works for our challenging right-rotationally invariant designs. A similar strategy may be invoked for any other penalty where an analogue of Theorem 1 is available. Therefore, we anticipate our approach to be broadly generalizable beyond ridge regression.

2 PRELIMINARIES AND SETUP

2.1 Right-rotationally invariant designs

Definition 1 (Right-rotationally invariant design)**.** A random design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is *right-rotationally invariant* if for any $\mathbf{O} \in \mathbb{O}(p)$, one has $\mathbf{XO} \overset{d}{=} \mathbf{X}$, where $\mathbb{O}(p)$ denotes $p \times p$ orthogonal matrices. Equivalently, $\mathbf{X}$ is right-rotationally invariant if and only if the singular value decomposition $\mathbf{X} = \mathbf{Q}^\top \mathbf{DO}$ satisfies that $\mathbf{O}$ is independent of $(\mathbf{Q}, \mathbf{D})$ and drawn from the Haar measure on $\mathbb{O}(p)$, i.e. "uniformly" distributed on $\mathbb{O}(p)$.

Some examples of right-rotationally invariant designs are as follows (note that an i.i.d. entries standard Gaussian design is a member of this class, but the class includes many non-i.i.d. and non-Gaussian designs as well; see Appendix 8.3 for proofs of invariance):

(i) Autocorrelated data: The rows of $\mathbf{X}$ can be drawn from an autocorrelated series, where $\mathbf{x}_i = \rho \mathbf{x}_{i-1} + \sqrt{1-\rho^2}\mathbf{z}_i$, with $\mathbf{z}_i$ being i.i.d. draws from $\mathsf{N}(0, \mathbf{I}_n)$.

(ii) t-distributed data: The rows of $\mathbf{X}$ can be drawn from a multivariate t distribution with as few as 3 degrees of freedom, capturing heavy-tailed data such as financial returns (Grobys, 2021).

(iii) Products of Gaussian matrices: $\mathbf{X} = \mathbf{X}_1\mathbf{X}_2 \cdots \cdots \mathbf{X}_k$, where $\mathbf{X}_1$ has n rows and $\mathbf{X}_k$ has p columns, while the remaining dimensions are arbitrary. See Hanin and Nica (2020); Hanin and Paouris (2021) for connections to linear neural networks.

(iv) Matrix-Normals: $\mathbf{X} \sim \mathsf{MN}(\mathbf{0}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$, where $\mathbf{\Sigma}^{\text{row}}$ can be arbitrary while $\mathbf{\Sigma}^{\text{col}} \sim \mathsf{InvWishart}(\mathbf{I}_p, (1+\delta)p)$, for any $\delta > 0$. See Appendix 8.5 for details on distribution notation.

(v) Equicorrelated data: $\mathbf{X} \in \mathbb{R}^{n\times p}$ has independent columns, but each column follows a multivariate Gaussian distribution with covariance matrix $\mathbf{\Sigma}$, where $\Sigma_{ij} = \rho$ if $i \neq j$, and $\Sigma_{ii} = 1$.

(vi) Spiked matrices: $\mathbf{X} = \lambda \mathbf{V}\mathbf{W}^\top + \mathbf{G}$, where $\mathbf{V} \in \mathbb{R}^{n\times r}$ and $\mathbf{W} \in \mathbb{R}^{p\times r}$ are the first r columns of two Haar matrices, and $G_{ij} \sim \mathsf{N}(0,1)$.

See (Li and Sur, 2023, Figure 1) for more such examples. Setting (ii) has heavier tails than supported by existing GCV theory; settings (i, iii, v, vi) have sample dependence; setting (iv) has both.

Right-rotationally invariant ensembles allow us enormous flexibility—they can simultaneously capture dependent rows and heavy-tails by allowing a rather general class of singular value distributions. The spectrum of right-rotationally invariant designs can essentially be arbitrary, and allowing for significant generalization from prior i.i.d. designs used to study the behavior of LOOCV and GCV in high dimensions. Furthermore, the universality class of these designs is extremely broad, as established in Dudeja et al. (2023); Wang et al. (2023).

To simplify some statements, we introduce some notation. Let $m_{\mathbf{D}}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{D_{ii}^2 - z}$ and $v_{\mathbf{D}}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{D_{ii}^2 - z}$, with $D_{ii} = 0$ if $i > \min(n,p)$. Also of use will be the derivatives $m'_{\mathbf{D}}$ and $v'_{\mathbf{D}}$. Note $m_{\mathbf{D}}$ is the Stieltjes transform of the empirical measure of $(\mathbf{D}^\top\mathbf{D}_{ii})_{i=1}^p$; $v_{\mathbf{D}}$ is the Stieltjes transform of $(\mathbf{D}\mathbf{D}_{ii}^\top)_{i=1}^n$.

2.2 Problem Setup

We study high-dimensional ridge regression and GCV with right-rotationally invariant designs. Concretely, we observe $(\mathbf{X}, \mathbf{y})$, such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

with $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{\epsilon}, \boldsymbol{\beta} \in \mathbb{R}^p$. Hence n is the sample size, p is both the data dimension and the parameter count, $\boldsymbol{\beta}$ is the signal, and $\boldsymbol{\epsilon}$ is a component-wise independent noise vector, where each entry has mean 0 and variance σ^2. We study the expected out-of-sample risk on an independent draw from a new, potentially different right-rotationally invariant design, $\widetilde{\mathbf{X}} = \widetilde{\mathbf{Q}}^\top \widetilde{\mathbf{D}} \widetilde{\mathbf{O}}$, with $\widetilde{\mathbf{X}} \in \mathbb{R}^{n'\times p}$. This corresponds to learning a model on one interdependent population, and using it on a separate interdependent population. We analyze the conditional risk

$$R_{\mathbf{X},\mathbf{y}}\left(\hat{\boldsymbol{\beta}}(\mathbf{X},\mathbf{y}), \boldsymbol{\beta}\right) = \frac{1}{n'}\mathbb{E}\left[\left\|\widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right\|^2 \middle| \mathbf{X}, \mathbf{y}\right]. \tag{2}$$

Such conditional risks have been examined in prior works (Hastie et al., 2022; Patil et al., 2021). We focus on the risk of the ridge estimator

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}, \tag{3}$$

which has closed form $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$. As such, sometimes we will write $R_{\mathbf{X},\mathbf{y}}(\lambda) := R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta})$ for convenience. Before stating our results, we note the running assumptions used in what follows.

- A1 We consider a sequence of right-rotationally invariant designs $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \ldots$ where each $\mathbf{X}_i \sim \mathbf{Q}_i^\top \mathbf{D}_i \mathbf{O}_i$, and a sequence of signal vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n, \ldots$ such that for all n, $\boldsymbol{\beta}_n \in \mathbb{R}^p$ and $\|\boldsymbol{\beta}_n\| = r\sqrt{n}$; $\mathbf{y}_i$ is generated by Eq. (1). In the sequel, we drop the dependence on n and write $\mathbf{X}, \mathbf{y}$ and $\boldsymbol{\beta}$ with the indexing implicit. We refer to $(\mathbf{X}, \mathbf{y})$ as the training data.
- A2 $\limsup \lambda_{\max}(\mathbf{D}_n) < C$ a.s. for some constant C.
- A3 $\mathbf{X}_n \in \mathbb{R}^{n\times p(n)}$, $n \to \infty$, $p(n)/n \to \gamma \in (0, \infty)$.
- A4 $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$; ϵ_i's are independent; $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \sigma^2$, and $\mathbb{E}[\epsilon_i^{4+\eta}] < \infty$ for some $\eta > 0$.
- A5 To fix scaling and simplify our results, we assume that $\frac{n\mathbb{E}[\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})]}{n'p} = 1$; this ensures the risk is defined and maintains the scale of the training set relative to the test set (See Remark 8 of Appendix).
- A6 $\boldsymbol{\beta}_n$ is fixed or random independent of $(\mathbf{X}_n, \boldsymbol{\epsilon})$.
- A7 $\widetilde{\mathbf{X}}$ is an independent draw from a potentially different right-rotationally invariant ensemble.

Of crucial importance will be the two parameters r^2 and σ^2, which dictate the behavior of the problem.

Remark 1. To clarify why the risk in Eq. 2 is relevant, if one samples $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ according to Eq. (1), then $R_{\mathbf{X},\mathbf{y}}(\lambda) = \mathbb{E}[\|\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_\lambda\|^2 \mid \mathbf{X}, \mathbf{y}] - \sigma^2$, meaning the risk defined corresponds, up to a factor of σ^2, to the mean-squared error on the test set.

Remark 2. One should think of $\mathbf{X}$ in our setting as akin to the normalized matrix $\mathbf{X}/\sqrt{n}$ in the i.i.d. setting. This further ensures that $\|\mathbf{X}^\top\mathbf{X}\|_{\mathsf{op}} = \|\mathbf{D}^\top\mathbf{D}\|_{\mathsf{op}} = \mathcal{O}(1)$. To maintain the signal to noise ratio under this scaling, $\|\boldsymbol{\beta}\| = \mathcal{O}(\sqrt{n})$ (Assumption A1); the test set is appropriately scaled through Assumption A5.

3 GENERALIZED CROSS-VALIDATION

Prior work (Xu et al., 2021; Patil et al., 2021) established that for many designs with i.i.d. samples, LOOCV and GCV estimate the out-of-sample error consistently in the proportional asymptotics regime, while k-fold CV does not. In our setting, LOOCV is intractable without additional assumptions on $\mathbf{Q}$ (see Appendix 8.6). However, GCV is tractable, motivating our study here. Initially introduced as a technique for selecting the regularization parameter in smoothing splines Craven and Wahba (1978), GCV's optimality and consistency has been shown in various tasks (Li, 1986; Gu and Ma, 2005; Zhang et al., 2015; Xu et al., 2018, 2019). The GCV cross-validation metric for ridge regression is as follows (Golub et al., 1979):

$$\mathsf{GCV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_\lambda}{1 - \mathsf{Tr}(\mathbf{S}_\lambda)/n}\right)^2. \tag{4}$$

Here $\mathbf{S}_\lambda = \mathbf{X}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top$ is the smoother matrix. Our first result is as follows:

Theorem 1. *Under the stated assumptions,*

$$\mathsf{GCV}_n(\lambda) - \frac{r^2(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) + \sigma^2\gamma v'_{\mathbf{D}}(-\lambda)}{\gamma v_{\mathbf{D}}(-\lambda)^2} \xrightarrow{a.s.} 0. \tag{5}$$

Further, the risk takes the following form:

Theorem 2. *Fix $0 < \lambda_1 < \lambda_2 < \infty$. Define*

$$\mathsf{R}_{\mathbf{D}}(r^2, \sigma^2) = r^2\left(\frac{\lambda^2}{\gamma}v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma}\right) + \sigma^2(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) \tag{6}$$

Under the stated assumptions,

$$\sup_{\lambda\in[\lambda_1,\lambda_2]} \left|R_{\mathbf{X},\mathbf{y}}\left(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}\right) - \mathsf{R}_{\mathbf{D}}(r^2, \sigma^2)\right| \xrightarrow{a.s.} 0. \tag{7}$$

where we recall that $r^2 = \|\boldsymbol{\beta}_n\|^2/n$ *and* $\sigma^2 = \mathbb{E}\epsilon_i^2$.

In Theorem 2, one can interpret the term involving r^2 as the bias of the estimator, and the term with σ^2 as its variance. As such, one can then interpret Theorem 1 as showing how GCV reconstructs the bias and variance from the data.

Remark 3. It is shown in Patil et al. (2021) that for i.i.d. designs, $\Delta_n(\lambda) := \mathsf{GCV}_n(\lambda) - R_{\mathbf{X},\mathbf{y}}(\lambda) - \sigma^2$ approaches zero uniformly over compact intervals[1]. Their proof of this relies explicitly on identities of the Stieltjes transform given by the Silverstein equation (Silverstein, 1995). These identities do not exist for right-rotationally invariant designs, and thus the GCV becomes biased for these designs, meaning $\Delta_n(\lambda) \not\to 0$. The more the spectrum departs from that of the i.i.d. setting, the more biased the estimates.

3.1 A Modified GCV

The original GCV (in Eq. (4)) can be seen as starting with the training error and finding a rescaling that recovers the out-of-sample risk. To construct our alternative GCV metric, which is provably consistent for right-rotationally invariant designs, we likewise begin with the training error, but instead use it to produce estimating equations for r^2 and σ^2. We then use these to compute the risk formula given by Theorem 2.

3.1.1 Estimators for r^2, σ^2

Consistent estimation of σ^2 has been studied in Li and Sur (2023). We propose a different approach since we observe it performs better in finite samples. In particular, we establish the following result:

Lemma 1. *Under our assumptions, for any* $\lambda > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda\right)^2 - \left[r^2\frac{\lambda^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)\right) + \sigma^2\lambda^2 v'_{\mathbf{D}}(-\lambda)\right] \xrightarrow{a.s.} 0. \quad (8)$$

Observe that Eq. (8) can be used as an estimating equation, meaning that, if one treats (8) as an equality to 0 (i.e. replacing "$\xrightarrow{a.s.}$" with "="), then we can solve for the unknown quantities r^2 and σ^2. Recall that $\gamma = p/n$, $v_{\mathbf{D}}$, and $v'_{\mathbf{D}}$, defined in Section 2.1, are all data dependent quantities. Now, Eq. (8) yields a distinct estimating equation for *each* value of λ. While it takes only two equations to solve for r^2 and σ^2, we instead use a grid of λ and find OLS estimates for r^2 and σ^2 to improve robustness. Explicitly, the scheme proceeds as shown in steps 1-6 of Algorithm 1.

[1]The extra σ^2 term is discussed in Remark 1.

Using Lemma 1, we can establish that the following:

Corollary 2. *Under Assumptions A1-A7 and a non-degeneracy condition on* $\mathbf{D}$ *(see Appendix 7.7.1 for details),* $\hat{r}^2$ *and* $\hat{\sigma}^2$ *from Algorithm 1 are strongly consistent for* r^2 *and* σ^2*, respectively.*

Empirically, we find that normalizing the spectrum of $\mathbf{X}^\top\mathbf{X}$ and taking $\{\lambda_i\}_{i=1}^L$ logarithmically spaced between 1 and $10^{2.5}$ works well.

Algorithm 1 ROTI-GCV(λ)

Input: $\mathbf{X}, \mathbf{y}$, λ; for use in estimation, grid of regularization strengths $\lambda_1 < \cdots < \lambda_L$;
Output: Estimate for $R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$

1: **for** $i = 1, \ldots, L$ **do**
2: $\quad a_i \leftarrow \frac{\lambda_i^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda_i) - \lambda_i v'_{\mathbf{D}}(-\lambda_i)\right)$
3: $\quad b_i \leftarrow \lambda_i^2 v'_{\mathbf{D}}(-\lambda_i)$
4: $\quad t_i \leftarrow \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda_i}\|_2^2$.
5: **end for**
6: $\hat{\sigma}^2_{\{\lambda_i\}_{i=1}^L}(\mathbf{X}, \mathbf{y}) \leftarrow \frac{\sum_{i=1}^L \left(\frac{t_i}{a_i} - \frac{t_1}{a_1}\right)\left(\frac{b_i}{a_i} - \frac{b_1}{a_1}\right)}{\sum_{i=1}^L \left(\frac{b_i}{a_i} - \frac{b_1}{a_1}\right)^2}$,
$\quad \hat{r}^2_{\{\lambda_i\}_{i=1}^L}(\mathbf{X}, \mathbf{y}) \leftarrow \frac{\sum_{i=1}^L \left(\frac{t_i}{b_i} - \frac{t_1}{b_1}\right)\left(\frac{a_i}{b_i} - \frac{a_1}{b_1}\right)}{\sum_{i=1}^L \left(\frac{a_i}{b_i} - \frac{a_1}{b_1}\right)^2}$.
7: **return** $\mathsf{R}_{\mathbf{D}}(\hat{r}^2, \hat{\sigma}^2)$ for $\mathsf{R}_{\mathbf{D}}(\cdot,\cdot)$ defined in (6)

3.1.2 Plug-in estimation

We now define $\mathsf{ROTI\text{-}GCV}(\lambda) := \mathsf{R}_{\mathbf{D}}(\hat{r}^2, \hat{\sigma}^2)$ (recall $\mathsf{R}_{\mathbf{D}}$ from Theorem 2) and tune λ using this metric (Algorithm 1); below we establish that this quantity is uniformly consistent for the out-of-sample risk over compact intervals.

Corollary 3. *Fix any* $0 < \lambda_1 < \lambda_2 < \infty$. *Under Assumptions A1-A7, one has*

$$\sup_{\lambda\in[\lambda_1,\lambda_2]}\left|\mathsf{ROTI\text{-}GCV}(\lambda) - R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta})\right| \xrightarrow{a.s.} 0. \quad (9)$$

Remark 4. Uniform convergence, unlike pointwise convergence, ensures the risk attained by the minimizer (over a compact interval) of our cross-validation metric is close to optimal. See Appendix 8.4 for a short proof.

4 GCV UNDER SIGNAL-PC ALIGNMENT

An issue that emerges when attempting to apply ROTI-GCV in practice is that the signal often tends to align with the top eigenvectors of the data, violating the independence assumption imposed on $\boldsymbol{\beta}$ and $\mathbf{O}$. The right-rotationally invariant assumption, for fixed $\boldsymbol{\beta}$, inherently assumes that the signal is incoherent with respect to the eigenbasis. However, this assumption is

generally not true, and in fact in the i.i.d. anisotropic case, the geometry of $\boldsymbol{\beta}$ with respect to the top eigenvectors of $\boldsymbol{\Sigma}$ is known to significantly influence the behavior of ridge regression (see e.g. Wu and Xu (2020)).

Adapting the approach given in Li and Sur (2023), we model this scenario using two index sets $\mathcal{J}_a$ and $\mathcal{J}_c$ of distinguished eigenvectors, which are *aligned* and *coupled* eigenvectors, respectively. We replace Assumptions A6 and A7 in Section 3 with the following:

A8 The true signal $\boldsymbol{\beta}$ is given as $\boldsymbol{\beta} = \boldsymbol{\beta}' + \sum_{i\in\mathcal{J}_a} \sqrt{n}\alpha_i \mathbf{o}_i$, where $\|\boldsymbol{\beta}'\|/\sqrt{n} = r$; $\boldsymbol{\beta}'$ is independent of $(\mathbf{X}, \boldsymbol{\epsilon})$. Here $\mathbf{o}_i$ refers to the ith row of $\mathbf{O}$, which is an eigenvector of $\mathbf{X}^\top\mathbf{X}$. An eigenvector $\mathbf{o}_i$ is an *aligned* eigenvector if $\mathbf{o}_i \in \mathcal{J}_a$, as then $\mathbf{o}_i$ aligns with the signal $\boldsymbol{\beta}$;

A9 The test data $\widetilde{\mathbf{X}} \sim \widetilde{\mathbf{Q}}^\top\widetilde{\mathbf{D}}\widetilde{\mathbf{O}}$ is drawn from a separate right-rotationally invariant ensemble; $\widetilde{\mathbf{O}}$ is distributed according to a Haar matrix with rows conditioned to satisfy $\mathbf{o}_i = \widetilde{\mathbf{o}}_i$ for all $i \in \mathcal{J}_c$. An eigenvector $\mathbf{o}_i$ is then *coupled* if $i \in \mathcal{J}_c$, as $\mathbf{o}_i$ is shared between the test and train sets.

We require two more assumptions:

A10 $\mathcal{J}_a, \mathcal{J}_c$, $\boldsymbol{\alpha} = (\alpha_i)_{i\in\mathcal{J}_a}$ are fixed, not changing with n, p; this ensures consistent estimation of these distinguished directions.

A11 $\limsup \|\mathbb{E}[\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}}]\|_{\mathsf{op}} < C$ for some constant C. This differs from Assumption A2, as it is on the test data.

The main difficulty that arises in our setting, compared to Li and Sur (2023), is that we must account for how the geometry of the eigenvectors of the test set relates to those of the training set. One should expect, for example, in some types of structured data that the top eigenvectors of the training data and test data are closely aligned. We account for this through the coupled eigenvectors condition (A9). Furthermore, without this coupling, any alignment between the eigenvectors of the training set and the signal $\boldsymbol{\beta}$ will not exist in the test set, since then the eigenvectors of the test set would be independent of this alignment. We are still interested in the same test risk $R_{\mathbf{X},\mathbf{y}}$ as in Eq. (2), except now note that $\widetilde{\mathbf{X}}$ and $\mathbf{X}$ are dependent.

Theorem 3. *Under Assumptions A1-A5 and A8-A11, for any $0 < \lambda_1 < \lambda_2 < \infty$, we have*

$$\sup_{\lambda\in[\lambda_1,\lambda_2]} \left| R_{\mathbf{X},\mathbf{y}}\left(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}\right) - \mathcal{R}_{\mathbf{X},\mathbf{y}}(r^2, \sigma^2, \boldsymbol{\alpha}) \right| \xrightarrow{a.s.} 0, \tag{10}$$

where

$$\mathcal{R}_{\mathbf{X},\mathbf{y}}(r^2, \sigma^2, \boldsymbol{\alpha}) = \mathcal{B}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}) + \mathcal{V}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}). \tag{11}$$

Let $\mathfrak{d}_i^2 = \mathbb{E}\left[\left(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}}\right)_{ii}\right]$ *and* $\mathfrak{d}_{\mathsf{bulk}}^2 = \frac{1}{p-|\mathcal{J}_c|}\sum_{i\notin\mathcal{J}_c} \mathfrak{d}_i^2$.

$$\mathcal{B}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}) = \frac{\lambda^2}{n}\sum_{i=1}^{p}\left[\frac{(r^2/\gamma + \mathbb{1}(i\in\mathcal{J}_a)n\alpha_i^2)}{(D_{ii}^2+\lambda)^2} \cdot \left(\mathfrak{d}_{\mathsf{bulk}}^2 + \mathbb{1}(i\in\mathcal{J}_c)(\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2)\right)\right]$$

$$\mathcal{V}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{D_{ii}^2\left(\mathfrak{d}_{\mathsf{bulk}}^2 + \mathbb{1}(i\in\mathcal{J}_c)(\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2)\right)}{(D_{ii}^2+\lambda)^2}$$

Remark 5. Comparing the two terms here with those in Theorem 1, observe that they are equal when $\mathcal{J}_a$ and $\mathcal{J}_c$ are empty. Furthermore, observe that each element of $\mathcal{J}_a$ adds a large contribution to the bias, scaling with the size of the alignment α_i; similarly, elements of $\mathcal{J}_c$ contribute additional bias and variance.

Remark 6. The assumption that the top eigenvectors of training and test sample covariance are exactly equal (Assumption A9) is not expected to hold perfectly in practice. However, the idea is that the approximation can lead to more robust procedures. We thus expect this approximation to hold when the training and test data have "spiky" spectra, where the top few eigenvalues are heavily separated from the bulk, and these top eigenvectors of the test and training data are close.

4.1 Estimating r^2, σ^2, and α_i

Our approach is simply to estimate the α_i's using the classical principal components regression (PCR); see Jolliffe (1982); Hubert and Verboven (2003) for details, also described in steps 1 and 2 of Algorithm 2. We then compute a transformed model which is right-rotationally invariant, from which Algorithm 1 steps 1-6 can be used to find r^2 and σ^2.

We require some notation; we focus on subsets S of the indices of all nonzero singular values $\mathcal{N} = \{i \in [p] : (\mathbf{D}^\top\mathbf{D})_{ii} > 0\}$. Let $\mathbf{O}_S \in \mathbb{R}^{|S|\times p}$ for any S denote the rows of $\mathbf{O}$ indexed by S. Let $\mathbf{D}_S \in \mathbb{R}^{(n-|S|)\times p}$ denote the rows of $\mathbf{D}$ indexed by S. Finally, for a set $S \subset \mathcal{N}$, denote $\overline{S} = \mathcal{N} \setminus S$. The estimation for r^2, σ^2, and $\boldsymbol{\alpha}$ then proceeds as in steps 1-5 of Algorithm 2.

Lemma 4. *Under the additional assumption that $\boldsymbol{\epsilon}$ has Gaussian entries*[2]*, $\tilde{r}^2$, $\tilde{\sigma}^2$, and $\hat{\boldsymbol{\alpha}}$, from Algorithm 2, are all strongly consistent.*

4.2 Cross-validation metric

Our approach for cross-validation is to then again explicitly compute the risk formulas $\mathcal{B}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta})$ and $\mathcal{V}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta})$, where we estimate the unknown parameters from data. To apply our method, we therefore

[2]The assumption of Gaussian entries is technical, and we believe it can be removed with some work.

need to consistently estimate $r^2, \sigma^2, \boldsymbol{\alpha}$, and $\{\mathfrak{d}_i^2\}_{i\in\mathcal{J}_c}$ and $\mathfrak{d}_{\mathsf{bulk}}^2$. Lemma 4 gives us most of these parameters, but we still require consistent estimates of the $\mathfrak{d}_i$'s. This is a nonissue in cases where $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ are modeled as deterministic. If they are random, we expect that if the test data is drawn from the same right-rotationally invariant ensemble as the training data, then, in practice, one can use the eigenvalues of the training data to do this. However, our framework further allows for the test distribution to be drawn from a different right-rotationally invariant ensemble. In such cases, it is possible to then use test data to do this estimation because it only depends on the test covariates $\widetilde{\mathbf{X}}$ (which the practitioner may have access to), but not its labels; this is done for experiments in Section 5. We therefore make one final assumption:

A11 We assume there exist estimators $\hat{\mathfrak{d}}_i$ and $\hat{\mathfrak{d}}_{\mathsf{bulk}}$ where we have $\sup_{i\in\mathcal{J}_a\cup\mathcal{J}_c}|\hat{\mathfrak{d}}_i - \mathfrak{d}_i| \xrightarrow{a.s.} 0$, and $|\hat{\mathfrak{d}}_{\mathsf{bulk}} - \mathfrak{d}_{\mathsf{bulk}}| \xrightarrow{a.s.} 0$.

We then take our validation metric as $\mathsf{aROTI\text{-}GCV}(\lambda) = \mathcal{R}_{\mathbf{X},\mathbf{y}}(\tilde{r}^2, \tilde{\sigma}^2, \hat{\boldsymbol{\alpha}}, \{\hat{\mathfrak{d}}_i\}_{i\in\mathcal{J}_c,\mathsf{bulk}})$, where the latter expression denotes substituting every parameter with its estimated version. Algorithm 2 summarizes our entire procedure[3].

Algorithm 2 aROTI-GCV(λ)

Input: $\mathbf{X}, \mathbf{y}$, regularization strength λ; index sets $\mathcal{J}_a, \mathcal{J}_c$; consistent estimators $\hat{\mathfrak{d}}_i^2$ and $\hat{\mathfrak{d}}_{\mathsf{bulk}}^2$; for use in estimation of r^2, σ^2, grid of regularization strengths $\lambda_1 < \cdots < \lambda_L$;

Output: Estimate for $R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$

1: $\mathbf{X}_{\mathsf{proj}} \leftarrow \mathbf{X}\mathbf{O}_{\mathcal{J}_a}^\top$
2: $\hat{\boldsymbol{\alpha}} \leftarrow (\mathbf{X}_{\mathsf{proj}}^\top\mathbf{X}_{\mathsf{proj}})^{-1}\mathbf{X}_{\mathsf{proj}}^\top\mathbf{y}/\sqrt{n}$
3: $\mathbf{y}_{\mathsf{new}} = (\mathbf{D}_{\bar{\mathcal{J}}_a}^\top\mathbf{D}_{\bar{\mathcal{J}}_a})^{1/2}(\mathbf{X}_{\mathsf{resid}}^\top\mathbf{X}_{\mathsf{resid}})^{-1}\mathbf{X}_{\mathsf{resid}}^\top\mathbf{y}$
4: $\mathbf{X}_{\mathsf{new}} = (\mathbf{D}_{\bar{\mathcal{J}}_a}^\top\mathbf{D}_{\bar{\mathcal{J}}_a})^{1/2}\mathbf{O}_{\bar{\mathcal{J}}_a}$
5: Compute $\tilde{r} = \hat{r}^2_{\{\lambda_i\}_{i=1}^L}(\mathbf{X}_{\mathsf{new}}, \mathbf{y}_{\mathsf{new}}), \tilde{\sigma}^2 = \hat{\sigma}^2_{\{\lambda_i\}_{i=1}^L}(\mathbf{X}_{\mathsf{new}}, \mathbf{y}_{\mathsf{new}})$
6: **return** $\mathcal{R}_{\mathbf{X},\mathbf{y}}\left(\tilde{r}^2, \tilde{\sigma}^2, \hat{\boldsymbol{\alpha}}, \{\hat{\mathfrak{d}}\}_{i\in\mathcal{J}_c,\mathsf{bulk}}\right)$

Corollary 5. *Under Assumptions A1-A5, A8-A11, and the additional assumption that the entries of $\boldsymbol{\epsilon}$ are Gaussian,* aROTI-GCV *satisfies*

$$\sup_{\lambda\in[\lambda_1,\lambda_2]} |\mathsf{aROTI\text{-}GCV}(\lambda) - R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta})| \xrightarrow{a.s.} 0. \quad (12)$$

Remark 7. This states that aROTI-GCV is uniformly consistent on compact intervals, ensuring the minimizer of aROTI-GCV attains risk that is close to optimal.

[3]The algorithm requires a choice of $\mathcal{J}_a, \mathcal{J}_c$; we give a brief overview of selecting $\mathcal{J}_a$ and $\mathcal{J}_c$ in Section 5.3, followed by a detailed discussion in Appendix 9.

5 NUMERICAL EXPERIMENTS

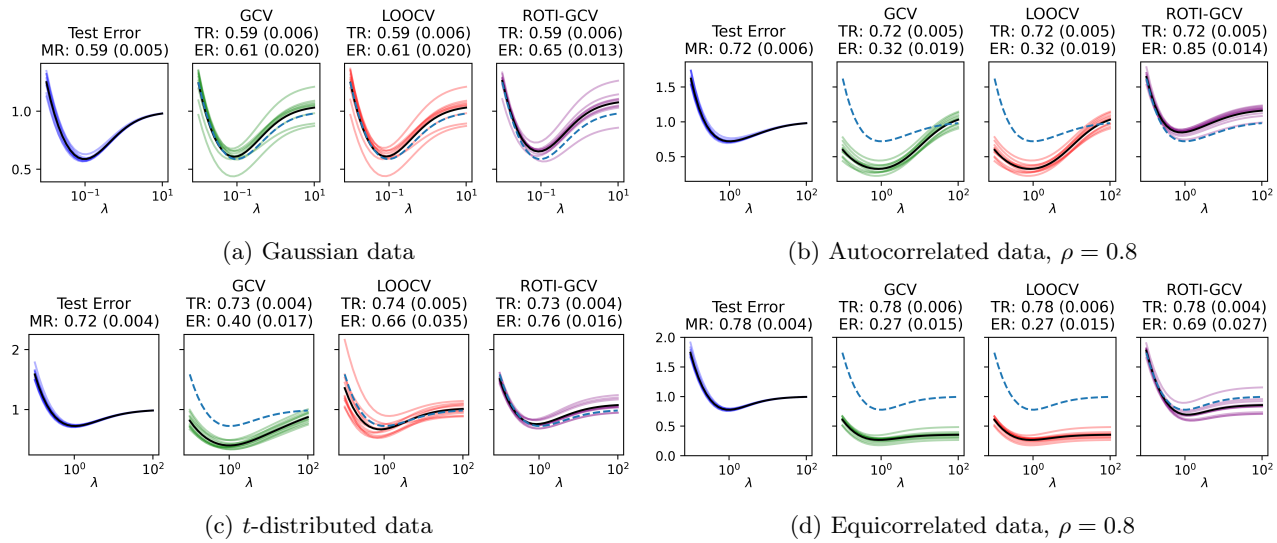
In this section, we demonstrate the efficacy of our method via numerical experiments. In producing each plot below, we generate a training set according to a prescribed distribution, as well as a signal vector $\boldsymbol{\beta}$. We then repeatedly resample the training noise $\boldsymbol{\epsilon}$ and run each cross-validation method, in addition to computing the true risk on the test set.

5.1 Experiments on right-rotationally invariant data

Figure 1 illustrates the performance of ROTI-GCV and compares it to LOOCV and GCV. In these figures, the Tuned Risk (abbreviated TR) refers to the out-of-sample risk obtained when we use the value of λ that optimizes the given cross-validation method. Minimum risk (MR), denotes the minimum of the true expected out-of-sample risk curve. Standard errors are shown in parentheses for all tuned risks. Estimated Risk (ER) refers to the value of the out-of-sample risk that would be estimated using the given cross-validation method. As expected, in the Gaussian setting all three methods perform well. However, for dependent data, such as in the autocorrelated and equicorrelated cases, we observe that the risk curves produced by GCV and LOOCV are wildly off from the truth, whereas ROTI-GCV accurately captures the true risk curve. Surprisingly, the tuned risk values from LOOCV, GCV are still decently close to that of ROTI-GCV. As discussed in Remark 3, GCV (and also LOOCV) estimate a quantity with an additional factor of σ^2; thus, to make the evaluation of estimated risks fair, these two curves are plotted with this term removed.

5.2 Experiments under Signal-PC alignment

We now consider settings where the signal $\boldsymbol{\beta}$ aligns with the eigenvectors $\mathbf{O}$. The relevant method here is aROTI-GCV (Algorithm 2). In Figure 2a, we show the performance of aROTI-GCV when we set the top eigenvectors of the test set equal to those of the training set, matching the coupled eigenvectors condition (Assumption A9). We further construct the signal such that it aligns with the top 10 eigenvectors. As the conditions match our theory, we see aROTI-GCV performs well. The setting of Figure 2a might look contrived, but we next provide examples of well-known distributions where such structures naturally arise (Figures 2b and 2c). In Figure 2b, the data is drawn from a mixture of two Gaussians, one centered at $\mathbf{3} = [3, 3, \ldots, 3]$ and the other centered at $-\mathbf{3}$, each with identity covariance. In this setting, the top eigenvector is very close to $\mathbf{3}$ and emerges from the bulk. Furthermore, projected into the subspace $\mathrm{span}(\mathbf{3})^\perp$, the



(a) Gaussian data

(b) Autocorrelated data, $\rho = 0.8$

(c) t-distributed data

(d) Equicorrelated data, $\rho = 0.8$

Figure 1: Risk curves produced by the cross-validation methods in addition to the true risk curve. Each plot uses a different form of right-rotationally invariant data, as indicated in each caption. (b): Autocorrelated data: rows of $\mathbf{X}$ are drawn according to $\mathbf{x}_i = \rho \mathbf{x}_{i-1} + \sqrt{1-\rho^2}\mathbf{z}_i$, with $\mathbf{z}_i$ being i.i.d. draws from $\mathsf{N}(0, \mathbf{I}_n)$. We set $\rho = 0.8$. (c): t-distributed data: each row is drawn from a multivariate t distribution with 3 degrees of freedom. (d): Equicorrelated data $\mathbf{X} \in \mathbb{R}^{n \times p}$ has independent columns, but each column follows a multivariate Gaussian distribution with covariance matrix $\mathbf{\Sigma}$, where $\Sigma_{ij} = \rho$ if $i \neq j$, and $\Sigma_{ii} = 1$. All simulations have $n = p = 1000$ and $r^2 = \sigma^2 = 1$. The x-axis for every plot is λ, the regularization parameter. The colored lines (blue, green, red, purple) are one of 10 iterations. In each iteration, we compute the cross-validation metric over a range of λ to produce the line, which reflects the estimated out-of-sample risk. The black line of each plot shows the average result for that method. The dashed blue line shows the average expected MSE curve as a benchmark.

data is a standard Gaussian and thus right-rotationally invariant. As seen, our method works under this type of covariance structure. Shown in Figure 2c, we find similar results when each row of $\mathbf{X}$ is drawn from an equicorrelated Gaussian, i.e., $\mathbf{x}_i \sim \mathsf{N}(0, \mathbf{\Sigma})$ where $\Sigma_{ij} = \rho^{1-\mathbb{1}(i=j)}$. Both examples mirror the setting of 2a in that they are not right-rotationally invariant, as the test and train tests vary in one direction more than the others. However, we see that the modifications in aROTI-GCV enable the method to still succeed. Note that in both of these examples, the top eigenvalue diverges at the rate $O(n)$, which violates our Assumption A11 as well as the bounded spectrum condition (Patil et al., 2021, Assumption 3), meaning neither cross-validation method is provably valid in this setting. However, we observe that ROTI-GCV still performs reasonably well.

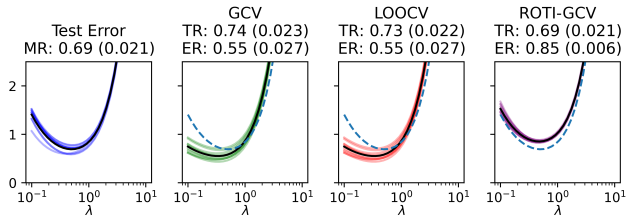
5.3 Experiments on semi-real data

In our semi-real experiments, we use real data for the designs, but generate the signals and responses ourselves. Knowing the true signal enables us to calculate the population risk of any estimator, allowing us to evaluate the performance of our GCV metrics accurately. For such experiments, we must verify whether the assumptions of our theory are met. To this end, we provide a series of diagnostics to check whether the most important assumptions hold, including to select the index sets $\mathcal{J}_a$ and $\mathcal{J}_c$. We briefly list these here, and defer more details to Appendix 9, including examples of running these diagnostics for multiple datasets.

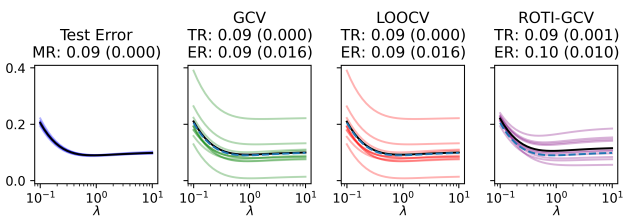
(i) regularity of singular value distributions (Assumptions A2, A11). Plot histogram of singular values to check for outliers; see Figure 3.

(ii) signal-PC alignment: relationship between $\boldsymbol{\beta}$ and $\{\mathbf{o}_i\}_{i=1}^p$ (Assumptions A6, A8). Li and Sur (2023) provide a hypothesis testing framework for identifying $\mathcal{J}_a$, wherein one computes a p-value per eigenvector to test whether or not it is aligned; the Benjamini-Hochberg procedure can then be used to control the false discovery rate. See Table 2.

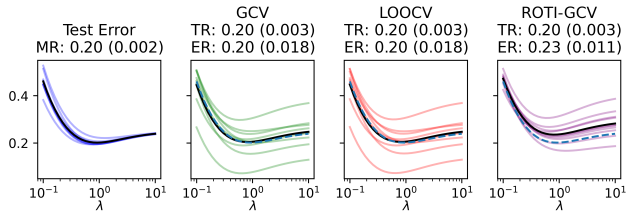
(iii) coupling: relationship between $\{\mathbf{o}_i\}_{i=1}^p$ and $\{\widetilde{\mathbf{o}}_j\}_{i=1}^p$ (Assumptions A7, A9). We compute overlaps $\sqrt{p}\langle \mathbf{o}_i, \widetilde{\mathbf{o}}_j \rangle$ for the top singular vectors. Note if $\mathbf{o}_i \perp\!\!\!\perp \widetilde{\mathbf{o}}_j$ (as under A7), then $\sqrt{p}\langle \mathbf{o}_i, \widetilde{\mathbf{o}}_j \rangle \to \mathsf{N}(0,1)$. If $\sqrt{p}\langle \mathbf{o}_i, \widetilde{\mathbf{o}}_j \rangle$ is large relative to this null distribution, $\mathbf{o}_i$, $\widetilde{\mathbf{o}}_j$ should be coupled. See Table 1. We leave development of a formal hypothesis test to future work.



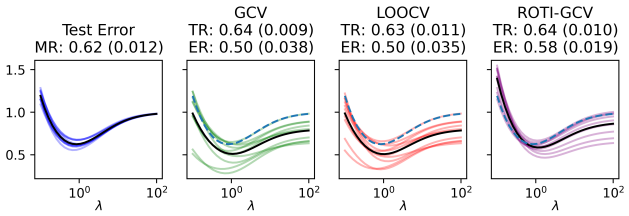
(a) Signal-PC alignment



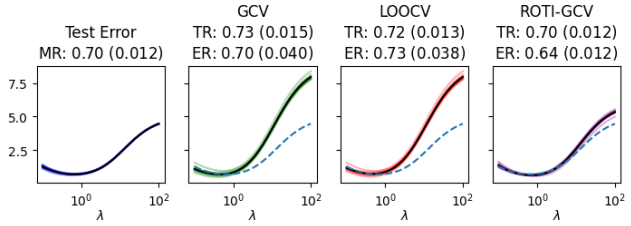
(b) Gaussian mixture



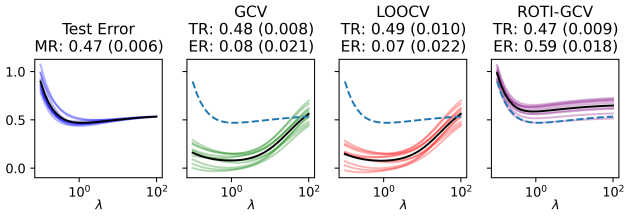
(c) Equicorrelated rows, $\rho = 0.5$



(d) Speech Data without Signal-PC alignment



(e) Speech Data with Signal-PC alignment



(f) Residualized returns

Figure 2: As in Figure 1, curves show the risk and estimated risk as a function of λ. (a): $\mathbf{X}$ has autocorrelated rows, with $\rho = 0.8$. The top 10 eigenvectors ($\mathcal{J}_c = [10]$) are coupled, so that the top 10 eigenvectors of the test set are equal to those of $\mathbf{X}$. $\boldsymbol{\beta} = \boldsymbol{\beta}' + \sqrt{n}\sum_{i=1}^{10} \frac{i}{10}\mathbf{o}_i$, with $\|\boldsymbol{\beta}'\|^2 = \sqrt{n}$, $\sigma^2 = 1$, $n = p = 1000$. (b) i.i.d. rows drawn from a Gaussian mixture, i.e. $\mathbf{x}_i$ is drawn from $\frac{1}{2}\mathsf{N}(\mathbf{3}, \mathbf{I}_p) + \frac{1}{2}\mathsf{N}(-\mathbf{3}, \mathbf{I}_p)$. When computing ROTI-GCV, we set $\mathcal{J}_c = \{1\}$. (c): i.i.d. rows drawn from an equicorrelated Gaussian, i.e. $\mathbf{x}_i \sim \mathsf{N}(0, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = \mathbb{1}(i = j) + \rho\mathbb{1}(i \neq j)$. We set $\rho = 0.5$. When computing ROTI-GCV, we set $\mathcal{J}_c = [10]$. (d): speech data with $\boldsymbol{\beta}$ sampled uniformly from sphere, with $r^2 = \sigma^2 = 1$, $n = p = 400$; we choose $\mathcal{J}_c = [3]$. (e): speech data with $\boldsymbol{\beta} = \boldsymbol{\beta}' + \frac{\sqrt{n}}{2}\sum_{i=1}^{5} \mathbf{o}_i$; again $r^2 = \sigma^2 = 1$ and $n = p = 400$. We choose $\mathcal{J}_c = [3]$ and $\mathcal{J}_a = [5]$ (discussion in Section 5.3.1). (f): 30 minute residualized returns sampled every 1 minute; $\boldsymbol{\beta}$ sampled uniformly from sphere.

5.3.1 Speech data

The designs we use consist of speech data sourced from OpenML (Le, 2017). Each row has dimension 400 and consists of an i-vector (a type of featurization for audio data, see e.g. Ibrahim and Ramli (2018)) of a speech segment. In this first example, we fully illustrate all the proposed diagonstics; as such, we consider both when $\boldsymbol{\beta}$ is uniformly drawn on the sphere, so there is no signal-PC alignment, and when signal-PC alignment is present, to illustrate how to choose $\mathcal{J}_a$.

We first examine the distribution of singular values in Figure 3, finding there are no outlier values. Next, we compute overlaps of test and train eigenvectors, shown in Table 1, finding that the top 3 eigenvectors of each have strong overlaps. This motivates setting $\mathcal{J}_c = [3]$[4].

The results for this setting when $\boldsymbol{\beta}$ is drawn uniformly from the sphere are shown in Figure 2d, where we see successfully capture the true loss curve and that the estimated risk values are superior to those given by GCV and LOOCV.

Next, in the setting with added signal-PC alignment, we instead take $\boldsymbol{\beta} = \boldsymbol{\beta}' + \frac{\sqrt{n}}{2}\sum_{i=1}^{5} \mathbf{o}_i$, meaning the aligned set is $\mathcal{J}_a = [5]$. We use the hypothesis testing framework of Li and Sur (2023) and compute Benjamini-Hochberg-adjusted p-values for alignment, shown in Table 2; here, we clearly identify the set of aligned eigenvectors. The resulting loss curves for this setting are then shown in Figure 2e.

5.3.2 Residualized returns

[4] One can further couple the next few eigenvectors (e.g. $\mathbf{o}_6$ with $\widetilde{\mathbf{o}}_4$ and so on), but we observe this has no impact on the results.

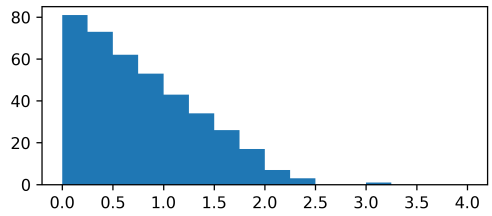


Figure 3: Singular value distribution for speech data.

Table 1: Overlaps for speech data setting; Cell (i, j) contains $\sqrt{p}|\langle \mathbf{o}_i, \widetilde{\mathbf{o}}_j \rangle|$.

	$\widetilde{\mathbf{o}}_1$	$\widetilde{\mathbf{o}}_2$	$\widetilde{\mathbf{o}}_3$	$\widetilde{\mathbf{o}}_4$	$\widetilde{\mathbf{o}}_5$	$\widetilde{\mathbf{o}}_6$	$\widetilde{\mathbf{o}}_7$	$\widetilde{\mathbf{o}}_8$	$\widetilde{\mathbf{o}}_9$	$\widetilde{\mathbf{o}}_{10}$
$\mathbf{o}_1$	17.30	2.72	0.63	2.01	1.55	1.02	0.30	0.96	1.70	0.27
$\mathbf{o}_2$	0.83	15.24	0.68	1.63	1.34	1.22	0.89	1.42	0.27	2.19
$\mathbf{o}_3$	1.08	1.93	10.93	1.60	0.64	2.02	4.63	0.35	4.14	2.75
$\mathbf{o}_4$	0.29	2.78	2.92	0.22	3.98	6.63	2.98	1.65	0.54	1.56
$\mathbf{o}_5$	1.27	2.70	4.02	2.14	5.11	1.29	4.34	4.72	2.95	2.24
$\mathbf{o}_6$	1.19	0.56	4.82	5.95	0.37	3.83	1.37	0.87	4.98	1.04
$\mathbf{o}_7$	2.21	0.70	2.14	2.65	4.84	1.77	0.54	1.00	1.65	2.53
$\mathbf{o}_8$	0.46	0.15	2.96	0.16	1.66	4.49	0.22	0.40	1.82	0.10
$\mathbf{o}_9$	0.30	0.40	0.72	0.01	1.63	1.39	3.86	4.53	1.12	2.57
$\mathbf{o}_{10}$	1.14	2.51	0.90	0.77	0.37	0.79	2.44	1.83	1.81	2.34

We next consider designs consisting of residualized returns from 493 stocks retrieved from Polygon API[5] (Polygon, 2024). There is one row per minute, and each contains the residualized return for each stock over the last 30 minutes. Since they share 29 minutes of returns, consecutive rows are heavily correlated. The regression corresponds to the known practice of replicating the residualized returns of an unknown asset $\mathbf{y}$ using a portfolio of stocks $\hat{\boldsymbol{\beta}}$ (with returns $\mathbf{X}\hat{\boldsymbol{\beta}}$). The diagnostic plots in this case are tame, and we require no coupling; these are deferred to Appendix 9. The loss curves produced by ROTI-GCV better capture the true loss curve, and that the estimated risk values (ER) provide superior estimates of the true out-of-sample risk compared to LOOCV and GCV.

6 DISCUSSION AND CONCLUSION

Takeaways: We give a precise analysis of ridge regression for right-rotationally invariant designs. These designs depart crucially from the existing framework of i.i.d. samples from distributions with light enough tails, instead allowing for structured dependence and heavy tails. We then characterize GCV in this setting, finding that it is inconsistent, and introduce ROTI-GCV, a consistent alternative; various synthetic and semi-real experiments illustrate the accuracy of our method. As a byproduct of our analysis, we produce new estimators for signal-to-noise ratio (r^2) and noise variance (σ^2) robust to these deviations from the i.i.d. assumptions. At the core, our method relies on noticing that the relation given by Lemma 1 yields a different estimating equation for every choice of λ; using this to generate a series of estimating equations, we obtain stable estimators of r^2, σ^2. This core idea applies whenever an analogue of Lemma 1 can be derived, giving it reach beyond ridge regression. As future steps, extending our method to other settings such as generalized linear models, kernel ridge regression, regularized random features regression would be important directions. Furthermore, recent advances in Dudeja et al. (2023); Wang et al. (2023) show that right-rotationally invariant distributions have a broad universality class – in fact as long as the singular vectors of a design matrix satisfy certain generic structure, properties of learning algorithms under such settings are well captured by results derived under the right-rotationally invariant assumption. In particular, the universality of the in-sample risk of ridge regression is already established in these works, and we believe such universality also holds for the out-of-sample risk; thus, our analysis should extend to the entire universality class of these designs, which is remarkably broad. This promises that our approach can have wide reach, allowing reliable hyperparameter tuning for a broad class of covariate distributions.

Limitations: The well-specification assumption (A1) is crucial and seems more difficult to relax for our designs than in the i.i.d. case due to the global dependence structure. Right-rotationally invariant designs have difficulty modeling distributions with feature covariances that are much more complicated than those covered in Section 5, especially when the spectrum is not dominated by eigenvalues. Finding ways to account for different types of feature covariance in data using these designs would naturally increase the broader applicability of our method, such as by modeling the design as $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}\boldsymbol{\Sigma}^{1/2}$ and extending our analysis. Finally, as mentioned in Section 1, our results focus on ridge regression, though we believe similar results for other penalties can be obtained.

Table 2: Benjamini-Hochberg adjusted p-values for alignment.

	$\mathbf{o}_1$	$\mathbf{o}_2$	$\mathbf{o}_3$	$\mathbf{o}_4$	$\mathbf{o}_5$
p	0.000	0.000	0.000	0.000	0.000
	$\mathbf{o}_6$	$\mathbf{o}_7$	$\mathbf{o}_8$	$\mathbf{o}_9$	$\mathbf{o}_{10}$
p	0.934	0.588	0.651	0.913	0.395

[5]We remove the top 8 right singular vectors from data matrix $\mathbf{X}$. This is analogous to *residualization* of returns, a common practice in financial modeling. We compute the factors to delete by taking the top PCs of the covariance matrix computed using both test and train data; PCA is crude/weak as a factor model and we have only 8 factors; hence it is more comparable to compute PCs looking at both train and test data. See Appendix 8.2 for more details on residualization.

References

Adamczak, R. (2015). A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(none):1 – 13.

Atanasov, A., Zavatone-Veth, J. A., and Pehlevan, C. (2024a). Risk and cross validation in ridge regression with correlated samples.

Atanasov, A. B., Zavatone-Veth, J. A., and Pehlevan, C. (2024b). Scaling and renormalization in high-dimensional regression.

Bigot, J., Dabo, I.-M., and Male, C. (2024). High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile. working paper or preprint.

Chazottes, J.-R. (2022). Notes on means, medians and gaussian tails. `https://hal.science/hal-03636138v1/file/Gaussian-concentration-around-the-mean-and-the-median.pdf`.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik*, 31(4):377–403.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.*, 46(1):247–279.

Dudeja, R., Sen, S., and Lu, Y. M. (2023). Spectral universality of regularized linear regression with nearly deterministic sensing matrices. *arXiv preprint arXiv:2208.02753*.

Fan, Z. (2022). Approximate Message Passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197 – 224.

Fan, Z. and Wu, Y. (2021). The replica-symmetric free energy for ising spin glasses with orthogonally invariant couplings. *arXiv preprint arXiv:2105.02797*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Gerbelot, C., Abbara, A., and Krzakala, F. (2020). Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1682–1713. PMLR.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Grobys, K. (2021). What do we know about the second moment of financial markets? *International Review of Financial Analysis*, 78:101891.

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics*, 33(3):1357 – 1379.

Hanin, B. and Nica, M. (2020). Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322.

Hanin, B. and Paouris, G. (2021). Non-asymptotic results for singular values of gaussian matrix products. *Geometric and Functional Analysis*, 31(2):268–324.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

Hubert, M. and Verboven, S. (2003). A robust pcr method for high-dimensional regressors. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(8-9):438–452.

Ibrahim, N. S. and Ramli, D. A. (2018). I-vector extraction for speaker recognition based on dimensionality reduction. *Procedia Computer Science*, 126:1534–1540.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):300–303.

Le, M.-A. (2017). Openml, dataset id: 40910. `https://www.openml.org/search?type=data&status=active&id=40910`. Accessed: 2024-10-10. This work is licensed under the CC-BY License.

Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, pages 1101–1112.

Li, Y., Fan, Z., Sen, S., and Wu, Y. (2023). Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Transactions on Information Theory*, 70(30):2118–2154.

Li, Y. and Sur, P. (2023). Spectrum-aware adjustment: A new debiasing framework with applications to principal component regression. *arXiv preprint arXiv:2309.07810*.

Liang, T. and Sur, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669–1695.

Liu, L., Huang, S., and Kurkoski, B. M. (2022). Memory amp. *IEEE Transactions on Information Theory*, 68(12):8015–8039.

Ma, J. and Ping, L. (2017). Orthogonal amp. *IEEE Access*, 5:2020–2033.

Mardia, K. V., Kent, J. T., and Taylor, C. C. (2024). *Multivariate analysis*, volume 88. John Wiley & Sons.

Meckes, E. S. (2019). *The Random Matrix Theory of the Classical Compact Groups*. Cambridge Tracts in Mathematics. Cambridge University Press.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pages 3178–3186. PMLR.

Polygon (2024). Polygon api. `https://polygon.io/docs/stocks/getting-started`. Accessed: 2024-10-10.

Rahnama Rad, K. and Maleki, A. (2018). A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82.

Rangan, S., Schniter, P., and Fletcher, A. K. (2019). Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684.

Schechtman, G. (2014). Concentration, results and applications. `https://www.weizmann.ac.il/math/gideon/sites/math.gideon/files/uploads/concentrationNov19_0.pdf`.

Silverstein, J. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339.

Song, Y., Lin, X., and Sur, P. (2024). Hede: Heritability estimation in high dimensions by ensembling debiased estimators. *arXiv preprint arXiv:2406.11184*.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175:487–558.

Tagliazucchi, E., von Wegner, F., Morzelewski, A., Brodbeck, V., Jahnke, K., and Laufs, H. (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. *Proceedings of the National Academy of Sciences*, 110(38):15419–15424.

Takeda, K., Uda, S., and Kabashima, Y. (2006). Analysis of cdma systems that are characterized by eigenvalue spectrum. *Europhysics Letters*, 76(6):1193.

Takeuchi, K. (2019). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386.

Wang, S., Zhou, W., Maleki, A., Lu, H., and Mirrokni, V. (2018). Approximate Leave-One-Out for High-Dimensional Non-Differentiable Learning Problems. *arXiv e-prints*, page arXiv:1810.02716.

Wang, T., Zhong, X., and Fan, Z. (2023). Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*.

Wu, D. and Xu, J. (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression.

Xu, G., Shang, Z., and Cheng, G. (2018). Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *International Conference on Machine Learning*, pages 5483–5491. PMLR.

Xu, G., Shang, Z., and Cheng, G. (2019). Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. *Journal of computational and graphical statistics*, 28(4):891–908.

Xu, J., Maleki, A., Rad, K. R., and Hsu, D. (2021). Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030.

Xu, Y., Hou, T., Liang, S., and Mondelli, M. (2022). Approximate message passing for multi-layer estimation in rotationally invariant models.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340.

Zscheischler, J., Naveau, P., Martius, O., Engelke, S., and Raible, C. C. (2021). Evaluating the dependence structure of compound precipitation and wind speed extremes. *Earth System Dynamics*, 12(1):1–16.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes; assumptions are listed clearly, as is the model;
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes. Algorithmic considerations are not the

focus of the paper and are deferred to supplement. The procedure is at most a constant times more expensive than computing a single ridge solution.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. Yes; All assumptions are stated clearly and marked as such; the algorithms are given explicitly;

   (b) Complete proofs of all theoretical results. Yes; located within supplemental information;

   (c) Clear explanations of any assumptions. Yes; remarks are given following assumptions

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, in supplemental material

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). yes, in supplemental material

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). yes, in supplemental material

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). yes, in supplemental material

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. Citation of OpenML dataset included.

   (b) The license information of the assets, if applicable. Included in citation entry.

   (c) New assets either in the supplemental material or as a URL, if applicable. Code included as URL in submission.

   (d) Information about consent from data providers/curators. License included in citation.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. not applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. not applicable

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. not applicable

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. not applicable

Appendix

7 Deferred Proofs

7.1 Proof of Lemma 6

Lemma 6. *Let $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}$ be right-rotationally invariant. Then $\mathbb{E}[\mathbf{X}^\top\mathbf{X}] = \mathsf{Tr}(\mathbb{E}[\mathbf{D}^\top\mathbf{D}])/p \cdot \mathbf{I}_p$.*

Proof. Let $\mathbf{O}$ have rows $\mathbf{o}_i$. Then

$$\mathbb{E}[\mathbf{X}^\top\mathbf{X}] = \mathbb{E}[\mathbf{O}^\top\mathbf{D}^\top\mathbf{D}\mathbf{O}] = \sum_{i=1}^{n\wedge p}\mathbb{E}[D_{ii}^2\mathbf{o}_i\mathbf{o}_i^\top] = \sum_{i=1}^{n\wedge p}\mathbb{E}[D_{ii}^2]\mathbb{E}[\mathbf{o}_i\mathbf{o}_i^\top]$$
$$\overset{(*)}{=} \frac{1}{p}\mathbf{I}_p\sum_{i=1}^{n\wedge p}\mathbb{E}[D_{ii}^2] = \mathsf{Tr}(\mathbb{E}[\mathbf{D}^\top\mathbf{D}])/p \cdot \mathbf{I}_p$$

where $(*)$ follows from $\mathbf{I}_p = \mathbb{E}[\mathbf{O}^\top\mathbf{O}] = \sum_{i=1}^p \mathbb{E}[\mathbf{o}_i\mathbf{o}_i^\top]$ and the exchangeability of the $\mathbf{o}_i$ (which is derived from the fact that $\mathbb{O}(p)$ contains the permutation matrices). □

7.2 Useful Results

Lemma 7 (Hanson-Wright for Spherical Vectors)**.** *Let $\mathbf{v} \in \mathbb{R}^n$ be a random vector distributed uniformly on S^{n-1} and let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a fixed matrix. Then*

$$\mathbb{P}(\left|\mathbf{v}^\top\mathbf{A}\mathbf{v} - \mathbb{E}[\mathbf{v}^\top\mathbf{A}\mathbf{v}]\right| \geq t) \leq 2\exp\left(-C\min\left(\frac{n^2t^2}{2K^4\|\mathbf{A}\|_F^2}, \frac{nt}{K^2\|\mathbf{A}\|}\right)\right).$$

for some absolute constants C, K.

Proof. Isoperimetric inequalities imply that vectors uniformly distributed on S^{n-1} have sub-Gaussian tails around their median (see Schechtman (2014)). Explicitly, it is stated that for $\mathbf{v} \sim \mathrm{Unif}(S^{n-1})$, one has that for any 1-Lispchitz function $f : \mathbb{R}^n \to \mathbb{R}$, one has

$$\mathbb{P}(|f(\mathbf{v}) - M| \geq t) \leq 2e^{-nt^2/2}, \tag{13}$$

where M denotes the median of $f(X)$ (i.e. $\min(\mathbb{P}(f(\mathbf{v}) \geq M), \mathbb{P}(f(\mathbf{v}) \leq M)) \geq \frac{1}{2}$). We now use (Chazottes, 2022, Theorem 1.2), which shows that having sub-Gaussian tails with respect to the median is equivalent to having sub-Gaussian tails with respect to the mean, with different constants. Thus, it holds that for any 1-Lipschitz function f, one has

$$\mathbb{P}(|f(\mathbf{v}) - \mathbb{E}f(\mathbf{v})| \geq t) \leq 4e^{-nt^2/32}. \tag{14}$$

Note that this bound is vacuous whenever $4e^{-nt^2/32} > 1$, and hence we can rewrite this as

$$\mathbb{P}(|f(\mathbf{v}) - \mathbb{E}f(\mathbf{v})| \geq t) \leq \min(4e^{-nt^2/32}, 1) \tag{15}$$

where we now note that for sufficiently large K, one has $\min(4e^{-nt^2/32}, 1) \leq 2e^{-nt^2/K}$ for any $n \geq 1$ and $t \geq 0$. Hence, one has

$$\mathbb{P}(|f(\mathbf{v}) - \mathbb{E}f(\mathbf{v})| \geq t) \leq 2e^{-nt^2/K}. \tag{16}$$

Finally, (Adamczak, 2015, Theorem 2.3) implies that random variables satisfying such types of concentration inequalities must in turn satisfy a Hanson-Wright type inequality. That is, Eq. (16) implies that, for fixed matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, one has

$$\mathbb{P}(\left|\mathbf{v}^\top\mathbf{A}\mathbf{v} - \mathbb{E}[\mathbf{v}^\top\mathbf{A}\mathbf{v}]\right| \geq t) \leq 2\exp\left(-C\min\left(\frac{n^2t^2}{2K^4\|\mathbf{A}\|_F^2}, \frac{nt}{K^2\|\mathbf{A}\|}\right)\right) \tag{17}$$

as desired. □

Lemma 8 (Expectations of Quadratic Forms)**.** *Let $\mathbf{v}$ be uniformly distributed on S^{n-1}, and let $\mathbf{A}$ be an $n \times n$ matrix. Then*

$$\mathbb{E}[\mathbf{v}^\top\mathbf{A}\mathbf{v}] = \mathsf{Tr}(\mathbf{A}\mathbb{E}[\mathbf{v}\mathbf{v}^\top]) = \frac{\mathsf{Tr}(\mathbf{A})}{n} \tag{18}$$

Proof. The first equality follows from the cyclic property of trace. The second can be seen to follow from the observation used in the proof of Lemma 6 that $\mathbb{E}[\mathbf{o}_i\mathbf{o}_i^\top] = \frac{1}{p}\mathbf{I}_p$ and the fact that columns of a Haar distributed matrix are uniform on the sphere (see e.g. Meckes (2019)). Alternatively, recall that for $\mathbf{g} \sim \mathsf{N}(0, \mathbf{I}_n)$, $\frac{\mathbf{g}}{\|\mathbf{g}\|} \perp\!\!\!\perp \|\mathbf{g}\|$ from (Mardia et al., 2024, Theorem 2.7.1). Hence

$$\mathbb{E}[\mathbf{v}\mathbf{v}^\top] = \mathbb{E}[\mathbf{g}\mathbf{g}^\top/\|\mathbf{g}\|^2] = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]/\mathbb{E}[\|\mathbf{g}\|^2] = \mathbf{I}_n/n.$$

□

Lemma 9. *Let $\mathbf{v}_n$ be a sequence of random vectors, where $\mathbf{v}_n$ is uniformly distributed on S^{n-1}. Let $\mathbf{u}_n$ be another sequence of random vectors independent of $\mathbf{v}_n$, where $\limsup \|\mathbf{u}_n\| < C$ for some constant C almost surely. Then $\mathbf{v}_n^\top \mathbf{u}_n \xrightarrow{a.s.} 0$.*

Proof. We replace $\mathbf{v}_n$ with $\frac{\mathbf{g}}{\|\mathbf{g}\|}$ where $\mathbf{g} \sim \mathsf{N}(0, \mathbf{I}_p/p)$. Then conditional on $\mathbf{u}_n$, we have $\mathbf{g}^\top \mathbf{u}_n \sim \mathsf{N}(0, \|\mathbf{u}_n\|^2/p)$. We then use Mill's inequality to show $\mathbf{g}^\top \mathbf{u}_n \xrightarrow{a.s.} 0$:

$$\mathbb{P}\left(|\mathbf{g}^\top \mathbf{u}_n| > t \mid \mathbf{u}_n\right) \le \sqrt{\frac{2}{\pi}} \frac{\|\mathbf{u}_n\|}{t \cdot \sqrt{p}} \exp\left(-t^2/2 \cdot \frac{p}{\|\mathbf{u}_n\|^2}\right).$$

Since $\limsup_{n\to\infty} \|\mathbf{u}_n\|$ is a.s. bounded, the bound above is summable in n, so by Borel-Cantelli, the convergence is almost sure. An application of Slutsky finishes, since we know $\|\mathbf{g}\| \xrightarrow{a.s.} 1$ by the strong Law of Large Numbers, and hence $\mathbf{g}^\top \mathbf{u}_n/\|\mathbf{g}\| \to 0$. □

7.3 Proof of Theorem 1

First recall the statement of Theorem 1: Under the stated assumptions,

$$\mathsf{GCV}_n(\lambda) - \frac{r^2(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) + \sigma^2\gamma v'_{\mathbf{D}}(-\lambda)}{\gamma v_{\mathbf{D}}(-\lambda)^2} \xrightarrow{a.s.} 0.$$

We analyze the numerator and denominator separately. The latter is simple:

$$(1 - \mathsf{Tr}(\mathbf{S}_\lambda)/n)^2 = \left(1 - \mathsf{Tr}((\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})/n\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n \frac{\lambda}{\lambda + D_{ii}^2}\right)^2 = \gamma^2\lambda^2 v_{\mathbf{D}}(-\lambda)^2$$

The numerator only requires slightly more work. Recall $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Furthermore, note that the GCV numerator can be rewritten as below:

$$\frac{1}{n}\mathbf{y}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{y} = \underbrace{\frac{1}{n}\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta}}_{T_1} + 2\underbrace{\frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta}}_{T_2} + \underbrace{\frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\boldsymbol{\epsilon}}_{T_3}.$$

We handle each term. First, T_1 simplifies easily into a quadratic form involving the uniformly random vector $\mathbf{O}\boldsymbol{\beta}$ which is independent of $\mathbf{D}$. This will later allow us to apply Lemma 7:

$$T_1 = \frac{1}{n}(\mathbf{O}\boldsymbol{\beta})^\top\mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D}(\mathbf{O}\boldsymbol{\beta})$$

The expectation of this quantity is as follows:

$$\begin{aligned}
\mathbb{E}[T_1] &= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\mathsf{Tr}(\mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D})}{p} \\
&= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{1}{p}\sum_{i=1}^p D_{ii}^2\left(1 - \frac{D_{ii}^2}{D_{ii}^2 + \lambda}\right)^2 \\
&= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\lambda^2}{p}\sum_{i=1}^p \frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2} \\
&= \frac{\|\boldsymbol{\beta}\|^2}{n}\lambda^2\left(m_{\mathbf{D}}(-\lambda) - \lambda m'_{\mathbf{D}}(-\lambda)\right) = \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\lambda^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)\right)
\end{aligned}$$

Now, letting $\mathbf{P} = \mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D}$, Hanson-Wright (Lemma 7) implies

$$\mathbb{P}(|T_1 - \mathbb{E}[T_1]| > t) \le 2\exp\left(-c\min\left(\frac{p^2t^2}{2K^4\|\mathbf{P}\|_F^2}, \frac{pt}{K^2\|\mathbf{P}\|}\right)\right) = 2\exp\left(p\min\left(\frac{t^2}{2K^4\frac{1}{p}\|\mathbf{P}\|_F^2}, \frac{t}{K^2\|\mathbf{P}\|}\right)\right)$$

Note that

$$\mathbf{P} = \operatorname{diag}\left(\frac{\lambda^2 D_{ii}^2}{(D_{ii}^2+\lambda)^2}\right)_{i=1}^p$$

Hence

$$\begin{aligned}\frac{1}{p}\|\mathbf{P}\|_F^2 &= \frac{1}{p}\sum_{i=1}^p\left(\frac{\lambda^2 D_{ii}^2}{(D_{ii}^2+\lambda)^2}\right)^2 = \frac{1}{p}\sum_{i=1}^p\left(\frac{\lambda^2}{(D_{ii}^2+\lambda)^2}\right)^2 D_{ii}^4 \le \|\mathbf{D}\|_{\mathsf{op}}^4\\ \|\mathbf{P}\|_{\mathsf{op}} &= \frac{\lambda^2 D_{11}^2}{(D_{11}^2+\lambda)^2} \le \|\mathbf{D}\|^2\end{aligned}$$

By assumption, both terms are almost surely bounded in the limit. Recalling that $p(n)/n \to \gamma$, the above bound is summable in n and hence Borel-Cantelli implies almost sure convergence.

To handle the second term, we use a slightly different method.

$$\begin{aligned}T_2 &= \frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta}\\ &= \frac{1}{n}(\mathbf{O}\boldsymbol{\beta})^\top\underbrace{\mathbf{D}^\top\mathbf{Q}(\mathbf{I} - \mathbf{S}_\lambda)^2\boldsymbol{\epsilon}}_{\mathbf{T}_4}\end{aligned}$$

Note that $\tilde{\boldsymbol{\beta}} = \mathbf{O}\boldsymbol{\beta}$ is a uniformly random vector of norm $\|\boldsymbol{\beta}\|$. We replace this with $\|\boldsymbol{\beta}\|\frac{\mathbf{g}}{\|\mathbf{g}\|}$ where $\mathbf{g} \sim \mathsf{N}(0, \mathbf{I}_p/p)$. Then conditional on $\mathbf{D}, \boldsymbol{\epsilon}$, we have $\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T}_4 \sim \mathsf{N}(0, \|\mathbf{T}_4\|/(pn))$. Furthermore note $\|\mathbf{T}_4\| \le \|\mathbf{D}^\top\|\|\mathbf{Q}\|\|\mathbf{I} - \mathbf{S}_\lambda\|^2\|\boldsymbol{\epsilon}\| \le \|\mathbf{D}^\top\|\|\boldsymbol{\epsilon}\|$. Some standard tail bounds complete this.

$$\frac{1}{n}\|\boldsymbol{\beta}\|\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\mathbf{T}_4 = \frac{\|\boldsymbol{\beta}\|}{\sqrt{n}}\frac{1}{\|\mathbf{g}\|}\left(\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T}_4\right) = r\cdot\frac{1}{\|\mathbf{g}\|}\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T_4}$$

Hence

$$\mathbb{P}\left(|\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T_4}| > t \mid \mathbf{D}, \boldsymbol{\epsilon}\right) \le \sqrt{\frac{2}{\pi}}\frac{\|\mathbf{T}_4\|}{t\cdot\sqrt{pn}}\exp\left(-t^2/2\cdot\frac{pn}{\|\mathbf{T}_4\|^2}\right).$$

We claim that $\limsup_{n\to\infty}\|\mathbf{T}_4\|/\sqrt{n}$ is bounded. This follows from the bound above, the law of large numbers applied to $\boldsymbol{\epsilon}$, and the assumption on $\mathbf{D}$. Hence the bound above is summable in n once more, so by Borel-Cantelli, the convergence is almost sure. An application of Slutsky finishes, since we know $\|\mathbf{g}\| \xrightarrow{a.s.} 1$ by the strong Law of Large Numbers. The convergence for T_3 follows directly from (Dobriban and Wager, 2018, Lemma C.3)

7.4 Proof of Theorem 2

Proof of Theorem 2. In order to understand how one should tune the regularization parameter in this problem, we first must understand the out of sample risk for a given value of λ. The risk admits a decomposition into

three terms:

$$\begin{aligned}\frac{1}{n'}\mathbb{E}\left[\left\|\widetilde{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\|_2^2 \mid \mathbf{X},\mathbf{y}\right] &= \frac{1}{n'}\cdot\frac{\mathbb{E}[\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})]}{p}\left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right\|_2^2 \\ &= \frac{\mathbb{E}[\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})]}{n'p}\|(\mathbf{I}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\|_2^2 \\ &= \frac{n\mathbb{E}[\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})]}{n'p}\Bigg[\underbrace{\frac{1}{n}\boldsymbol{\beta}(\mathbf{I}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}}_{T_1} \\ &\quad + \underbrace{\frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-2}\mathbf{X}^\top\boldsymbol{\epsilon}}_{T_2} - \underbrace{2\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}}_{T_3}\Bigg]\end{aligned}$$

Remark 8. The normalization assumption (Assumption A5) ensures that the prefactor is finite. Without further loss of generality, we can assume it is 1. In fact it is sufficient for all proofs to simply assume that $\limsup \frac{n\mathbb{E}[\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})]}{n'p} \le C$, but assuming it is constant simplifies notation.

For intuition why this term is bounded, note that $\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}}) = \mathsf{Tr}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})$, which is the covariance matrix of n' samples normalized by $\sqrt{n}$ (the reason for this is given in Remark 2 – since $\boldsymbol{\beta}$ is scaled up by $\sqrt{n}$ all data must be scaled down by $\sqrt{n}$). Hence we expect this trace to be order $n'p/n$, which exactly cancels.

The first term corresponds to bias, the second to the variance, and the third is negligible. For the first term,

$$\begin{aligned}T_1 &= n^{-1}\boldsymbol{\beta}(\mathbf{I}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta} \\ &= n^{-1}(\mathbf{O}\boldsymbol{\beta})^\top\left[\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D}\right]^2(\mathbf{O}\boldsymbol{\beta}) \\ &= n^{-1}(\mathbf{O}\boldsymbol{\beta})^\top\left(\lambda(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\right)^2(\mathbf{O}\boldsymbol{\beta})\end{aligned}$$

Again let $\mathbf{b} = (\mathbf{O}\boldsymbol{\beta})/\|\boldsymbol{\beta}\|$ and $\mathbf{P} = \left(\lambda(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\right)^2$, and note $\mathsf{Tr}(\mathbf{P}) = \sum_{i=1}^p \frac{\lambda^2}{(D_{ii}^2+\lambda)^2}$ and $\|\mathbf{P}\|_F^2 = \sum_{i=1}^p \frac{\lambda^4}{(D_{ii}^2+\lambda)^4}$. Furthermore, one has $\mathbb{E}[\mathbf{b}^\top\mathbf{P}\mathbf{b}] = \frac{1}{p}\mathsf{Tr}(\mathbf{P})$ using 8, and clearly $\|\mathbf{P}\| \le 1$. We now apply Lemma 7 to show concentration:

$$\begin{aligned}\mathbb{P}\left(\left|\mathbf{b}^\top\mathbf{P}\mathbf{b}-\frac{1}{p}\mathsf{Tr}(\mathbf{P})\right| \ge t\right) &\le 2\exp\left(-C\min\left(\frac{p^2t^2}{2K^4\|\mathbf{P}\|_F^2}, \frac{pt}{K^2}\right)\right) \\ &= 2\exp\left(-Cp\min\left(\frac{t^2}{2K^4\frac{1}{p}\|\mathbf{P}\|_F^2}, \frac{t}{K^2}\right)\right)\end{aligned}$$

Note that $\frac{1}{p}\|\mathbf{P}\|_F^2$ is always bounded above by 1, so this bound can never be made vacuous by a certain setting of $\mathbf{D}$. This bound is summable in n and thus we have almost sure convergence of the difference to zero.

This yields pointwise convergence for the bias. To strengthen this to uniform convergence, we show that the derivative of the difference $\mathbf{b}^\top\mathbf{P}\mathbf{b} - \frac{1}{p}\mathsf{Tr}(\mathbf{P})$ is almost surely bounded on $[\lambda_1, \lambda_2]$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbf{b}^\top\mathbf{P}\mathbf{b} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{n}\sum_{i=1}^p(\mathbf{o}_i^\top\boldsymbol{\beta})^2\frac{\lambda^2}{(D_{ii}^2+\lambda)^2}\right] &= \frac{1}{n}\sum_{i=1}^p(\mathbf{o}_i^\top\boldsymbol{\beta})^2\frac{2\lambda(D_{ii}^2+\lambda)^2+2(D_{ii}^2+\lambda)(\lambda^2)}{(D_{ii}^2+\lambda)^4} \\ &\le \frac{1}{n}\frac{2\lambda(D_{11}^2+\lambda)^2+2(D_{11}^2+\lambda)(\lambda^2)}{(\lambda)^4}\sum_{i=1}^p(\mathbf{o}_i^\top\boldsymbol{\beta})^2 \\ &= \frac{1}{n}\frac{2\lambda(D_{11}^2+\lambda)^2+2(D_{11}^2+\lambda)(\lambda^2)}{(\lambda)^4}\end{aligned}$$

which is indeed almost surely bounded on $[\lambda_1, \lambda_2]$ in the limit by assumption on $\|\mathbf{D}\|$. Similarly, $\frac{\mathrm{d}}{\mathrm{d}\lambda}(\frac{1}{p}\mathsf{Tr}(\mathbf{P}))$ is also bounded on this interval by the same quantity, meaning the difference is Lipschitz. Hence one can discretize

the interval $[\lambda_1, \lambda_2]$ into a sufficiently fine grid. Pointwise convergence on every point holds at every point on the grid, then the Lipschitz condition assures that the convergence is uniform.

For the third term T_3, we rewrite it as

$$2(\boldsymbol{\epsilon}/\sqrt{n})^\top \mathbf{Q}\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{D} \cdot (\mathbf{O}\boldsymbol{\beta}/\sqrt{n})$$

and then apply Lemma 9 to see that this converges almost surely to zero. We can likewise control the derivative uniformly over the interval and again obtain almost-sure convergence.

Lastly, for the second term T_2, we first compute

$$\mathbb{E}\left[\frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^\top\boldsymbol{\epsilon}\right] = \frac{\sigma^2}{n}\sum_{i=1}^{n\wedge p}\frac{D_{ii}^2}{(D_{ii}^2+\lambda)^2}. \tag{19}$$

Then by (Dobriban and Wager, 2018, Lemma C.3), this term satisfies

$$\mathbb{E}\left[\frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^\top\boldsymbol{\epsilon}\right] - \frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^\top\boldsymbol{\epsilon} \xrightarrow{a.s.} 0. \tag{20}$$

The derivative is again uniformly controlled, producing uniform convergence once more.

□

7.5 Proof of Corollary 3

Proof. Under the assumptions of Theorem 1, for any consistent estimators $(\hat{r}^2, \hat{\sigma}^2)$ of (r^2, σ^2), respectively, one has, for any $\lambda_0 > 0$,

$$\sup_{\lambda>\lambda_0}\left|(\hat{r}^2 - r^2)\left(\frac{\lambda^2}{\gamma}v'_{\mathbf{D}}(-\lambda) + \frac{\gamma-1}{\gamma}\right) + (\hat{\sigma}^2 - \sigma^2)(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda))\right| \xrightarrow{p} 0. \tag{21}$$

If $\hat{r}^2, \hat{\sigma}^2$ are strongly consistent, then naturally the convergence is almost sure. This is immediate from the fact that

$$\lambda^2 v'_{\mathbf{D}}(-\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^2}{(D_{ii}^2+\lambda)^2} \le 1$$
$$v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) = \frac{1}{n}\sum_{i=1}^{n\wedge p}\frac{D_{ii}^2}{(D_{ii}^2+\lambda)^2} \le \frac{1}{4\lambda}.$$

Combining Eq. (21) with Theorem 2 then completes the corollary. □

7.6 Proof of Lemma 1

This was proven in the course of Theorem 1, as the training error is nothing but the numerator of the original GCV.

7.7 Proof of Corollary 2

7.7.1 Nondegeneracy on the spectrum of D

The exact nondegeneracy condition we require is composed of two portions:

1. Define the events $A_n = \bigcup_{i=1}^{L}\{a_i = 0\}$ and $B_n = \bigcup_{i=1}^{L}\{b_i = 0\}$. Then $\mathbb{1}(A_n \cup B_n) \xrightarrow{a.s.} 0$.
2. There exists $i_* \in \{2, \dots, L\}$ such that $\liminf |a_{i_*}b_1 - a_1 b_{i_*}| > C$ almost surely for some $C > 0$.

The need for the first condition is clear: our estimator would not even be defined if the entire spectrum were zero; in fact, the problem would be impossible, as the entire data matrix would be zero. The second is more complex, and we give a detailed discussion on its necessity before giving an explicit proof.

Note furthermore that each of these terms can be observed when using the procedure - this gives the user a good way to check if this assumption is likely satisfied in practice.

As motivation, we first discuss the simpler case when $L = 2$, as this has connections to the assumptions required for consistent estimation in Li and Sur (2023). In this setting, the estimator for σ^2 becomes

$$\frac{\left(\frac{t_2}{a_2} - \frac{t_1}{a_1}\right)\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)}{\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)^2} = \frac{\left(\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)\sigma^2 + o(1) \cdot \left(\frac{1}{a_2} - \frac{1}{a_1}\right)\right)\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)}{\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)^2}$$
$$= \sigma^2 + o(1)\frac{a_1 - a_2}{a_1 b_2 - b_1 a_2}.$$

where the first line comes from Lemma 1. Similarly, the estimator for r^2 is then $r^2 + o(1)\frac{b_2 - b_1}{a_1 b_2 - b_1 a_2}$. From this and the fact that the a_i and b_i are all bounded above, we see that this error term is asymptotically negligible provided $a_1 b_2 - b_1 a_2$ is bounded away from zero in the limit. In the limiting case where $\lambda_2 = \infty$, one has $a_2 = \frac{1}{p}\sum_{i=1}^p D_{ii}^2$ and $b_2 = 1$, meaning we require

$$\liminf |a_1 b_2 - a_2 b_1| = \liminf \left| \frac{1}{p}\sum_{i=1}^p \frac{\lambda^2 D_{ii}^2}{(D_{ii}^2 + \lambda)^2} - \left(\frac{1}{p}\sum_{i=1}^p D_{ii}^2\right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\lambda^2}{(D_{ii}^2 + \lambda)^2}\right)\right| > c > 0 \tag{22}$$

almost surely for some c. This is exactly the second part of our nondegeneracy assumption. Furthermore, this is the exact analog of (Li and Sur, 2023, Assumption 7) in our setting. The above argument thus establishes that consistent estimation is possible if one assumes equation (22).

We do not believe this condition to be simply technical. Consider the setting where $n \leq p$ and every $D_{ii} = 1$ for $i \leq n$ (meaning $D_{ii} = 0$ for $i > n$). Let $\tilde{\boldsymbol{\beta}} = \mathbf{O}\boldsymbol{\beta}$. Then one is given $\mathbf{y} = \mathbf{Q}^\top \tilde{\boldsymbol{\beta}}_{[1:n]} + \boldsymbol{\epsilon}$ and the goal is to infer $r^2 = \|\boldsymbol{\beta}\|_2^2/\sqrt{n} = \|\tilde{\boldsymbol{\beta}}\|_2^2/\sqrt{n}$ and σ^2. Every entry y_i is now $\mathbf{q}_i^\top \tilde{\boldsymbol{\beta}}_{[1:n]} + \epsilon_i$; the first term has mean zero and variance approximately r^2, and the second has mean zero and variance σ^2, and it essentially becomes impossible to decouple the effects of r^2 and σ^2. In this setting, the term above is exactly zero, reflecting that the estimating equations become linearly dependent. That said, we believe this issue should not affect practical use cases; the term we require to be bounded away from zero can be computed by the practitioner directly, and while it is impossible to check if some quantity is bounded away in the limit, they can observe its magnitude.

7.7.2 Proof

Having now overviewed a specific case of our method that yields consistent estimation through the use of an additional assumption already used in the literature, we now turn to the general setting. Note that in the general setting, the condition is weaker – we only require that $\liminf |a_1 b_i - a_i b_1| > C > 0$ holds for *one* of the i, meaning one loses nothing by using multiple λ_i, and in fact in practice we observe this results in improved stability.

Here the estimators take on the following forms:

$$\hat{\sigma}^2 = \sigma^2 + o(1)\frac{\sum_{i=1}^L (a_i^{-1} - a_1^{-1})(b_i a_i^{-1} - b_1 a_1^{-1})}{\sum_{i=1}^L (b_i a_i^{-1} - b_1 a_1^{-1})^2}$$
$$\hat{r}^2 = r^2 + o(1)\frac{\sum_{i=1}^L (b_i^{-1} - b_1^{-1})(a_i b_i^{-1} - a_1 b_1^{-1})}{\sum_{i=1}^L (a_i b_i^{-1} - a_1 b_1^{-1})^2}.$$

As a result, we will require that the coefficient of both error terms is uniformly bounded, which will yield that our estimators are consistent, meaning it suffices to show

$$\limsup \max\left(\left|\frac{\sum_{i=1}^L (a_i^{-1} - a_1^{-1})(b_i a_i^{-1} - b_1 a_1^{-1})}{\sum_{i=1}^L (b_i a_i^{-1} - b_1 a_1^{-1})^2}\right|, \left|\frac{\sum_{i=1}^L (b_i^{-1} - b_1^{-1})(a_i b_i^{-1} - a_1 b_1^{-1})}{\sum_{i=1}^L (a_i b_i^{-1} - a_1 b_1^{-1})^2}\right|\right) < C \tag{23}$$

almost surely for some C.

Now, note that

$$a(\lambda) = \frac{\lambda^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)\right) = \frac{1}{p}\sum_{i=1}^{p} D_{ii}^2\left(1 - \frac{D_{ii}^2}{D_{ii}^2+\lambda}\right)^2$$
$$b(\lambda) = \lambda^2 v'_{\mathbf{D}}(-\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^2}{(D_{ii}^2+\lambda)^2}$$

are both increasing in λ, and thus a_1 is the smallest of the a_i; likewise for b_1. Furthermore, note that the a_i are bounded above by $\|\mathbf{D}\|_{\mathsf{op}}$ which is then bounded almost surely by assumption; the b_i are bounded above by 1.

We then calculate as follows:

$$\begin{aligned}\frac{\left|\sum_{i=1}^{L}(a_i^{-1}-a_1^{-1})(b_ia_i^{-1}-b_1a_1^{-1})\right|}{\sum_{i=1}^{L}(b_ia_i^{-1}-b_1a_1^{-1})^2} &\le \frac{\left|\sum_{i=1}^{L}(a_i^{-1}-a_1^{-1})(b_ia_i^{-1}-b_1a_1^{-1})\right|}{(b_{i_*}a_{i_*}^{-1}-b_1a_1^{-1})^2}\cdot\frac{(a_1a_{i_*})^2}{(a_1a_{i_*})^2}\\ &= \frac{\left|\sum_{i=1}^{L}(a_{i_*}(a_1a_i^{-1})-a_{i_*})(b_ia_{i_*}(a_1a_i^{-1})-b_1a_{i_*})\right|}{(b_ia_1-b_1a_{i_*})^2}\\ &\le \frac{1}{(b_ia_1-b_1a_{i_*})^2}\sum_{i=1}^{L}\left|(a_{i_*}(a_1a_i^{-1})-a_{i_*})\right|\left|(b_ia_{i_*}(a_1a_i^{-1})-b_1a_{i_*})\right|\end{aligned}$$

Note now that since a_1 is the smallest of the a_i, $a_1a_i^{-1} \le 1$ for all i; hence, combined with the fact that the a_i and b_i are all bounded above in the limit, the final term is also controlled in the limit.

$$\limsup \frac{\left|\sum_{i=1}^{L}(a_i^{-1}-a_1^{-1})(b_ia_i^{-1}-b_1a_1^{-1})\right|}{\sum_{i=1}^{L}(b_ia_i^{-1}-b_1a_1^{-1})^2} \overset{a.s.}{\le} \frac{1}{C'}L\cdot C''\cdot 2\cdot C''$$

where C' is the constant from the non-degeneracy assumption and C'' is the constant from Assumption A2.

The same computation can be performed for the other term in (23) which proceeds exactly as above, but with the a_i's and b_i's exchanged. The result is

$$\limsup \frac{\left|\sum_{i=1}^{L}(b_i^{-1}-b_1^{-1})(a_ib_i^{-1}-a_1b_1^{-1})\right|}{\sum_{i=1}^{L}(a_ib_i^{-1}-a_1b_1^{-1})^2} \overset{a.s.}{\le} \frac{1}{C'}L\cdot 2\cdot C'',$$

as desired.

7.8 Proof of Theorem 3

Proof. Most computations are analogous to the proof of Theorem 2. We first compute $\mathbb{E}[\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} \mid \mathbf{X}, \mathbf{y}]$ using (Rangan et al., 2019, Lemma 4). We find that it is equal to $\sum_{i\in\mathcal{J}_c}(\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2)\mathbf{o}_i\mathbf{o}_i^\top + \mathfrak{d}_{\mathsf{bulk}}^2\mathbf{I}$. Call this quantity $\mathbf{C}$ for convenience. We then move to computing the risk.

$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} = (\mathbf{I} - (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} - (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}. \tag{24}$$

Meanwhile

$$\mathbb{E}\left[\|\widetilde{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2\right] = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\left(\sum_{i\in\mathcal{J}_c}(\mathfrak{d}_i^2-\mathfrak{d}_{\mathsf{bulk}}^2)\mathbf{o}_i\mathbf{o}_i^\top + \mathfrak{d}_{\mathsf{bulk}}^2\mathbf{I}\right)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}). \tag{25}$$

Thus as before, there are three main terms to handle.

Variance term:

$$\begin{aligned}&\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{C}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\\ &= \mathfrak{d}_{\mathsf{bulk}}^2\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon}+\\ &\qquad\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\left(\sum_{i\in\mathcal{J}_c}(\mathfrak{d}_{ii}^2-\mathfrak{d}_{\mathsf{bulk}}^2)\mathbf{e}_i\mathbf{e}_i^\top\right)(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon}.\end{aligned}$$

Note now that

$$\frac{n}{n'}\mathfrak{d}_{\mathsf{bulk}}^2 = \frac{n}{n'}\frac{1}{p-|\mathcal{J}_c|}\sum_{i\notin\mathcal{J}_c}\mathfrak{d}_i^2 = \frac{n\mathbb{E}\mathsf{Tr}(\widetilde{\mathbf{D}}^\top\widetilde{\mathbf{D}})}{n'p} + O(1/p) = 1 + O(1/p), \tag{26}$$

where the penultimate equality is due to assumption A11 and the last is due to the scaling assumption A5. Note that the $O(1/p)$ term converges almost surely to zero.

Then

$$\begin{aligned}
&\frac{n}{n'}\mathfrak{d}_{\mathsf{bulk}}^2\left|\frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon} - \frac{1}{n}\mathbb{E}[\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon}]\right| \\
&\leq |\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon} - \mathbb{E}[\boldsymbol{\epsilon}^\top\mathbf{Q}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{Q}\boldsymbol{\epsilon}]| + O(1/p) \xrightarrow{a.s.} 0.
\end{aligned}$$

The same arguments as in Theorem 2 then imply uniform convergence to zero.

On the other hand, the second term is order 1 and hence vanishes when divided by n. The convergence is then uniform because the function is monotonic in λ. It is almost sure through, for instance, using (Dobriban and Wager, 2018, Lemma C.3). Note that the form of $\mathcal{V}_{\mathbf{X}}$ contains an extra

$$\frac{\sigma^2}{n}\sum_{i\in\mathcal{J}_c}\frac{D_{ii}^2}{D_{ii}^2+\lambda}\mathfrak{d}_i^2$$

term which is uniformly negligible by monotonicity and the assumption that $\mathfrak{d}_i^2$ is a.s. bounded in the limit.

Bias term:

$$\begin{aligned}
&\boldsymbol{\beta}^\top(\mathbf{I}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})^\top\mathbf{C}(\mathbf{I}-(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}))\boldsymbol{\beta} \\
&= \mathfrak{d}_{\mathsf{bulk}}^2(\mathbf{O}\boldsymbol{\beta})^\top(\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})^2(\mathbf{O}\boldsymbol{\beta}) \\
&\quad + (\mathbf{O}\boldsymbol{\beta})^\top\left[(\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1})\left[\sum_{i\in\mathcal{J}_c}(\mathfrak{d}_i^2-\mathfrak{d}_{\mathsf{bulk}}^2)\mathbf{e}_i\mathbf{e}_i^\top\right](\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\right](\mathbf{O}\boldsymbol{\beta}).
\end{aligned}$$

We now need to handle this a bit more carefully than before, because $\boldsymbol{\beta}$ is no longer independent of $\mathbf{O}$. Note that $\mathbf{O}\boldsymbol{\beta} = \mathbf{O}\boldsymbol{\beta}' + \sum_{i\in\mathcal{J}_a}\alpha_i\mathbf{e}_i$. There are a few different types of terms we need to handle.

1. Unaligned terms:

$$\begin{aligned}
&\mathfrak{d}_{\mathsf{bulk}}^2(\mathbf{O}\boldsymbol{\beta}')^\top(\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})^2(\mathbf{O}\boldsymbol{\beta}') \\
&\quad + (\mathbf{O}\boldsymbol{\beta})^\top\left[(\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})\left[\sum_{i\in\mathcal{J}_c}(\mathfrak{d}_i^2-\mathfrak{d}_{\mathsf{bulk}}^2)\mathbf{e}_i\mathbf{e}_i^\top\right](\mathbf{I}-(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})\right](\mathbf{O}\boldsymbol{\beta})
\end{aligned}$$

   The first term is treated equivalently to the bias in the ridge case in Theorem 2, where, as in the variance computation above, the $\mathfrak{d}_{\mathsf{bulk}}^2$ can be handled. This produces uniform convergence to its expectation. The second term is $O(|\mathcal{J}_c|)$ and hence vanishes when divided by n. It does so uniformly because it is monotone increasing in λ.

2. Aligned but uncoupled terms:

$$\sum_{i\in\mathcal{J}_a\setminus\mathcal{J}_c} n\alpha_i^2\mathfrak{d}_{\mathsf{bulk}}^2\left(1-\frac{D_{ii}^2}{D_{ii}^2+\lambda}\right)$$

3. Aligned and coupled terms:

$$\sum_{i\in\mathcal{J}_a\cap\mathcal{J}_c} n\alpha_i^2\mathfrak{d}_i^2\left(1-\frac{D_{ii}^2}{D_{ii}^2+\lambda}\right)$$

4. Cross terms:

$$\sum_{i\in\mathcal{J}_a} \mathbf{o}_i^\top \boldsymbol{\beta}' \sqrt{n}\alpha_i \left[(\mathfrak{d}_{\mathsf{bulk}}^2 + \mathbb{1}(i\in\mathcal{J}_c)(\mathfrak{d}_{ii}^2 - \mathfrak{d}_{\mathsf{bulk}}^2)) \left(1 - \frac{D_{ii}^2}{D_{ii}^2+\lambda}\right)\right]$$

These terms are only of order $O(\sqrt{n})$ and hence vanish - they do so uniformly because, after taking absolute value of the summand, the resulting term is monotone increasing in λ, and thus it suffices to control the value at λ_2.

Note that the cross terms between $\alpha_i \mathbf{e}_i$ and $\alpha_j \mathbf{e}_j$ are all identically zero and thus do not need to be considered.

Summing all of these together, one obtains uniform convergence to $\mathcal{B}$. Note that the statement of $\mathcal{B}$ contains an extra

$$\frac{\lambda^2}{n}\sum_{i\in\mathcal{J}_c} \frac{r^2}{\gamma}\frac{\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2}{(D_{ii}^2+\lambda)^2} \tag{27}$$

term, which is uniformly negligible.

Cross term:

$$\begin{aligned}
&\boldsymbol{\epsilon}^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{C}(\mathbf{I} - (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} \\
&= \boldsymbol{\epsilon}^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{C}(\mathbf{I} - (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})(\boldsymbol{\beta}' + \sum_{i\in\mathcal{J}_a}\alpha_i\mathbf{o}_i) \\
&= \boldsymbol{\epsilon}^\top \mathbf{Q}\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{O}\mathbf{C}\mathbf{O}^\top(\mathbf{I} - \mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D})\mathbf{O}\boldsymbol{\beta}' + \\
&\qquad \sum_{i\in\mathcal{J}_a\cap\mathcal{J}_c} \mathbf{q}_i^\top\boldsymbol{\epsilon}\frac{D_{ii}}{\lambda+D_{ii}^2}\cdot\mathfrak{d}_i^2\cdot\frac{\lambda}{D_{ii}^2+\lambda}\cdot\sqrt{n}\alpha_i
\end{aligned}$$

For the first term, note that $\mathbf{O}\mathbf{C}\mathbf{O}^\top$ is in fact independent of $\mathbf{O}$. Moreover, the matrix in the middle has bounded operator norm, and thus this term vanishes after dividing by $1/n$ due to Lemma 9, where we apply the strong law of large numbers to bound $\|\boldsymbol{\epsilon}\|$. For the second, we note that this is almost surely, for sufficiently large n, uniformly bounded above by

$$C\sqrt{n}\cdot(\max_{i\in\mathcal{J}_a}\alpha_i)\cdot\max_{i\in\mathcal{J}_a\cap\mathcal{J}_c}|\mathbf{q}_i^\top\boldsymbol{\epsilon}|$$

Recall we are dividing everything by n. Hence if we show that $|\mathbf{q}_i^\top\boldsymbol{\epsilon}|/\sqrt{n} \xrightarrow{a.s.} 0$, then we will be done with a union bound. We rewrite $|\mathbf{q}_i^\top\boldsymbol{\epsilon}| = \boldsymbol{\epsilon}^\top(\mathbf{q}_i\mathbf{q}_i^\top)\boldsymbol{\epsilon}$ and can now apply (Dobriban and Wager, 2018, Lemma C.3) to find

$$\frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{q}_i\mathbf{q}_i^\top)\boldsymbol{\epsilon} - \frac{\sigma^2}{n} \xrightarrow{a.s.} 0.$$

Thus $|\mathbf{q}_i^\top\boldsymbol{\epsilon}|^2/n$ converges a.s. to 0, and thus by continuous mapping so does $|\mathbf{q}_i^\top\boldsymbol{\epsilon}|/\sqrt{n}$.

□

7.9 Proof of Lemma 4

The following is taken from Li and Sur (2023). We note that in this work, the errors are assumed Gaussian. We will mark which ones require this and how we believe they can be relaxed. First, the consistency of α_i is immediate from the first result of (Li and Sur, 2023, (40)), where we note that the proof does not require Gaussianity of the errors. The one step in the proof where care must be made is in controlling

$$\frac{1}{p}\left\|\mathbf{O}_{\mathcal{J}}^\top\left(\mathbf{D}_{\mathcal{J}}^\top\mathbf{D}_{\mathcal{J}}\right)^{-1}\mathbf{D}_{\mathcal{J}}^\top\mathbf{Q}\varepsilon\right\|_2^2, \tag{28}$$

but we note that this can be done through (Dobriban and Wager, 2018, Lemma C.3).

Then the reduced model is right-rotationally invariant by (Li and Sur, 2023, p. 58, (124)). Showing that the reduced model is right-rotationally invariant requires Gaussianity and is what leads to the added assumption of Lemma 4. Without the Gaussian assumption, the errors in the reduced model will not be i.i.d. However, in general we only require that certain inner products concentrate. We believe this will still hold in the reduced model when one explicitly tracks the form of the error.

7.10 Proof of Corollary 5

We will prove uniform convergence term by term. We begin with the variance. Define the following:

$$A = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{D_{ii}^2}{(D_{ii}^2+\lambda)^2}\left(\mathfrak{d}_{\mathsf{bulk}}^2 + \mathbb{1}(i \in \mathcal{J}_c)(\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2)\right)\right]$$

$$\hat{A} = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{D_{ii}^2}{(D_{ii}^2+\lambda)^2}\left(\hat{\mathfrak{d}}_{\mathsf{bulk}}^2 + \mathbb{1}(i \in \mathcal{J}_c)(\hat{\mathfrak{d}}_i^2 - \hat{\mathfrak{d}}_{\mathsf{bulk}}^2)\right)\right]$$

$$T_1 = \sigma^2 A \qquad T_2 = \hat{\sigma}^2 \hat{A} \qquad T_3 = \hat{\sigma}^2 A$$

Note that $T_1 = \mathcal{V}_{\mathbf{X}}$. First, note that $\hat{A}$ and $\hat{\sigma}^2/n$ are both almost surely bounded for sufficiently large n. This follows from the fact that the estimators are strongly consistent for bounded quantities. We can thus uniformly control $T_1 - T_3$ and then $T_3 - T_2$, producing uniform convergence.

For the bias, we wish to uniformly control

$$\mathcal{B}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}) = \frac{\lambda^2}{n}\sum_{i=1}^{p} \frac{\left(r^2/\gamma + \mathbb{1}(i \in \mathcal{J}_a) n\alpha_i^2\right)\left(\mathfrak{d}_{\mathsf{bulk}}^2 + \mathbb{1}(i \in \mathcal{J}_c)(\mathfrak{d}_i^2 - \mathfrak{d}_{\mathsf{bulk}}^2)\right)}{(D_{ii}^2+\lambda)^2}.$$

One can then expand the numerator into 4 constituent terms. Arguments similar to that of the variance then show uniform convergence for each of them.

8 Miscellaneous Details

8.1 Experimental details

All experiments were run on a personal laptop. Each experiment took only on the order of 1 minute. All experiments are carried out with $n = p = 1000$, because this scenario is most sensitive to the amount of regularization needed (as with $\lambda = 0$, often times the risk is infinite). In Section 3, we first sample a training set. We then repeatedly resample the training noise $\boldsymbol{\epsilon}$ because this is more economical. Each time, we then run each cross-validation technique and plot the loss curve over λ. The estimated risk is then the risk reported by the loss curve at its minimum. The tuned risk is then the actual mean-squared error (which has an analytical expression) produced when using the estimator with that value of λ. For ROTI-GCV, Assumption A5, while useful theoretically, proves to provide some practical challenges, mainly that finding the correct rescaling of the data that makes this assumption hold is not always simple, since the trace does not necessarily concentrate. In the experiments in Section 3, the data is drawn according to distributions satisfying this assumption, and the data is not normalized after being sampled. That said, normalizing the data does not affect any results. It is important to note that this scaling never affects the tuned value of λ and instead only affects the risk estimate.

In Section 4, it becomes necessary to estimate the eigenvalues of the test set data. To do this, we now sample a test set ($n = p = 1000$ again), and then compute the eigenvalues of the test set. A similar result should hold when using eigenvalues of the training set, since here both sets are equal in distribution.

8.2 Semi-real experimental details

8.2.1 Speech data

Speech data was taken directly from Le (2017) and then was standardized (demeaned and rescaled to have variance 1), as is usual when using ridge-regression.

8.2.2 Residualized returns

1. There is one row in X for each minute. Each row is the residualized return over the last 30 minutes, so consecutive rows are heavily correlated, since they share 29 minutes of returns. X has 493 rows and 493 columns ($n = p = 493$).
2. We generate β uniformly on the sphere of radius $r^2\sqrt{n}$, and the training labels $y = X\beta + \epsilon$, where $\epsilon_i \sim N(0, \sigma^2)$.
3. We plot the predicted loss curve given by each cross-validation metric
4. We compute the actual loss curve over the test data.

8.3 Proof of right-rotational invariance

Both the equicorrelated and autocorrelated cases can be written as $\mathbf{\Sigma}^{1/2}\mathbf{G}$, where $\mathbf{G}$ has i.i.d. standard Gaussian entries; the rotational invariance of the Gaussian then implies the right-rotational variance of $\mathbf{\Sigma}^{1/2}\mathbf{G}$ in these cases.

t-distributed data: A multivariate t distribution with ν degrees of freedom can be generated by first sampling $\mathbf{y} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$ and $u \sim \chi_\nu$; then $\mathbf{x} = \mathbf{y}/\sqrt{u/\nu}$; the rotational invariance of the Gaussian implies that this distribution is rotationally invariant. This proves rotational invariance for a single sample; stacking multiple independent samples into a matrix retains the rotational invariance.

Products of Gaussian matrices: follows from rotational invariance of the Gaussian.

Matrix normals: Recall that $\mathbf{\Sigma}^{\text{col}}$ is drawn from an inverse Wishart distribution with identity scale and $(1+\delta)p$ degrees of freedom; this can be generated by $(\mathbf{G}_2^\top \mathbf{G}_2)^{-1}$, where $\mathbf{G}_2 \in \mathbb{R}^{(1+\delta)p \times p}$ has i.i.d. standard Gaussian entries (since the inverse Wishart has identity scale) (Gelman et al., 2004, Appendix A). A matrix normal with row covariance $\mathbf{\Sigma}^{\text{row}}$ and column covariance $\mathbf{\Sigma}^{\text{col}} = (\mathbf{G}_2^\top \mathbf{G}_2)^{-1}$ can then be generated by $(\mathbf{\Sigma}^{\text{row}})^{1/2}\mathbf{G}(\mathbf{G}_2^\top \mathbf{G}_2)^{-1/2}$ (with $\mathbf{G}$ again having i.i.d. Gaussian entries and being independent of $\mathbf{G}_2$). The claim then follows from the following. For any fixed $\mathbf{O} \in \mathbb{O}_p$,

$$\begin{aligned}(\mathbf{\Sigma}^{\text{row}})^{1/2}\mathbf{G}(\mathbf{G}_2^\top \mathbf{G}_2)^{-1/2}\mathbf{O} &\overset{d}{=} (\mathbf{\Sigma}^{\text{row}})^{1/2}\mathbf{G}\mathbf{O}(\mathbf{O}\mathbf{G}_2^\top \mathbf{G}_2\mathbf{O}^\top)^{-1/2}\mathbf{O} \\ &= (\mathbf{\Sigma}^{\text{row}})^{1/2}\mathbf{G}(\mathbf{G}_2^\top \mathbf{G}_2)^{-1/2}\end{aligned}$$

as required.

Spiked matrices: It suffices to show that for any fixed $\mathbf{O}$ that $(\lambda\mathbf{V}\mathbf{W}^\top, \mathbf{G}) \overset{d}{=} (\lambda\mathbf{V}\mathbf{W}^\top\mathbf{O}, \mathbf{G}\mathbf{O})$. Since the first consists of an independent pair, it suffices to show that $\mathbf{V}\mathbf{W}^\top \overset{d}{=} \mathbf{V}\mathbf{W}^\top\mathbf{O}$, $\mathbf{G} \overset{d}{=} \mathbf{G}\mathbf{O}$ and that $\lambda\mathbf{V}\mathbf{W}^\top\mathbf{O}$ is independent of $\mathbf{G}\mathbf{O}$. The last two statements are automatic from the rotational invariance of the Gaussian and the independence of the $\mathbf{V}\mathbf{W}^\top$ and $\mathbf{G}$. The first can be seen from the fact that $\mathbf{W} = \mathbf{W}'\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{p\times r}$ has $\mathbf{P}_{ij} = \mathbb{1}(i=j)$ and $\mathbf{W}' \in \mathbb{R}^{p\times p}$ is Haar. Then $\mathbf{V}\mathbf{W}^\top\mathbf{O} = \mathbf{V}\mathbf{P}^\top(\mathbf{W}')^\top\mathbf{O} \overset{d}{=} \mathbf{V}\mathbf{P}^\top(\mathbf{W}')^\top = \mathbf{V}\mathbf{W}^\top$ where the distributional equality follows from the rotational invariance of Haar matrices.

8.4 Uniform convergence

Suppose one has a set of (possibly random) functions f_n satisfy, for some deterministic f,

$$\sup_{x\in[x_1,x_2]} |f_n(x) - f(x)| \xrightarrow{a.s.} 0. \tag{29}$$

Define

$$x_{\mathsf{cv},n} = \operatorname*{arg\,min}_{x\in[x_1,x_2]} f_n(x) \qquad x_* = \operatorname*{arg\,min}_{x\in[x_1,x_2]} f(x).$$

Then

$$|f(x_{\mathsf{cv},n}) - f(x_*)| \to 0. \tag{30}$$

Here the random functions f represent cross validation metrics which uniformly converge to the true asymptotic out-of-sample risk f. This result then shows that the minimizer chosen by the cross-validation metric tends towards the asymptotically optimal value. Note that this result cannot be obtained with only pointwise convergence.

Proof. Note $f(x_{\mathsf{cv},n}) - f(x_*) \geq 0$ always. Hence it suffices to show an upper bound which tends to zero.

$$\begin{aligned}f(x_{\mathsf{cv},n}) - f(x_*) &= f(x_{\mathsf{cv},n}) - f_n(x_{\mathsf{cv},n}) + \underbrace{f_n(x_{\mathsf{cv},n}) - f_n(x_*)}_{\leq 0} + f_n(x_*) - f(x_*) \\ &\leq |f(x_{\mathsf{cv},n}) - f_n(x_{\mathsf{cv},n})| + |f_n(x_*) - f(x_*)| \\ &\leq 2 \sup_{x\in[x_1,x_2]} |f_n(x) - f(x)| \to 0.\end{aligned}$$

as required. □

8.5 Distribution notation

- $\mathsf{InvWishart}(\mathbf{\Psi}, \nu)$ denotes an inverse Wishart distribution (Gelman et al., 2004, Appendix A) with scale matrix $\mathbf{\Psi}$ and degrees of freedom ν.
- $\mathsf{MN}(\boldsymbol{\mu}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$ denotes a matrix normal with among-row covariance $\mathbf{\Sigma}^{\text{row}}$ and among-column covariance $\mathbf{\Sigma}^{\text{col}}$.

8.6 LOOCV Intractability

The LOOCV in our setting is difficult to analyze because the objective depends on $\mathbf{Q}$, on which we do not place any assumptions. In particular, one can derive

$$\begin{aligned}\mathsf{LOOCV}_n(\lambda) &= \mathbf{y}^\top(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{D}_\lambda^{-2}(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{y}/n \\ &= \mathbf{y}^\top(\mathbf{I}-\mathbf{Q}^\top(\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{Q})\mathbf{D}_\lambda^{-2}(\mathbf{I}-\mathbf{Q}^\top(\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{Q})\mathbf{y}/n\end{aligned}$$

where $\mathbf{D}_\lambda = \text{diag}((1-(\mathbf{S}_\lambda)_{ii})_{i=1}^n)$. We try to simplify the diagonal term first.

$$\begin{aligned}(D_\lambda)_{ii} &= 1-\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{x}_i \\ &= 1-\mathbf{q}_i^\top\mathbf{D}\mathbf{O}(\mathbf{O}^\top\mathbf{D}^\top\mathbf{D}\mathbf{O}+\lambda\mathbf{I})^{-1}\mathbf{O}^\top\mathbf{D}^\top\mathbf{q}_i \\ &= 1-\mathbf{q}_i^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{q}_i\end{aligned}$$

where $\mathbf{q}_i$ is the i-th row of $\mathbf{Q}$. We analyze

$$\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top = \sum_{i=1}^n \mathbf{q}_i\mathbf{q}_i^\top(\mathbf{D}_\lambda^{-2})_{ii}.$$

Then LOOCV should contribute two terms, as the cross term should vanish. They are

$$(\mathbf{O}\boldsymbol{\beta})^\top\mathbf{D}^\top\mathbf{Q}(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top)(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{O}\boldsymbol{\beta}$$

which should concentrate around its trace.

The second is

$$\boldsymbol{\epsilon}^\top(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top)(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\boldsymbol{\epsilon}$$

which also concentrates around some multiple of its trace.

Finding this trace now involves numerous terms with $\mathbf{Q}$. Everything except the center term is diagonal, but immediately one arrives at the following difficulty:

$$\begin{aligned}(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q})_{ii} &= \sum_{j=1}^n(\mathbf{q}_j\mathbf{q}_j^\top)_{ii}(\mathbf{D}_\lambda^{-2})_{jj} \\ &= \sum_{j=1}^n(\mathbf{q}_{ji})^2(\mathbf{D}_\lambda^{-2})_{jj}\end{aligned}$$

Assuming that the rows of $\mathbf{Q}$ are exchangeable conditional on $\mathbf{D}$ is insufficient to make progress. If $\mathbf{Q}$ is assumed to also be Haar, then likely closed forms can be derived, but this is not pursued.

8.7 Algorithm Complexity

Assuming matrix multiplications and inversions cost $O(p^3)$, ROTI-GCV takes $O(p^3)$ time. The entire process takes only a constant factor more time than computing a single ridge solution, as the fitting process implicitly requires fitting multiple ridge solutions.

9 Verifying assumptions for real data

Here, we work through applications of our method. We take a look at 5 application scenarios. Our focus will be on detecting $\mathcal{J}_c$, since that is the new concept introduced in this work. For each of the first four examples, we will generate the signal vector uniformly on the sphere, so there is no signal alignment. Afterwards, we will discuss a 5th example with signal-PC alignment. We refer to (Li and Sur, 2023, Section 4) for more detailed discussion on alignment detection, and the details on hypothesis testing for signal-PC alignment, presenting a method for selecting $\mathcal{J}_a$ that controls the false discovery rate. **For each, we discuss how to check whether the most important assumptions hold.** This includes how the index sets $\mathcal{J}_a$ and $\mathcal{J}_c$ (aligned/coupled eigenvectors, definitions both restated below) can be selected, and how our estimation scheme performs.

The five examples we discuss are as follows:

1. residualized returns, where each row contains the residualized returns a 30 minute interval; this was shown in Figure 2f, and are analyzed in more detail in Figure 4.
2. the gaussian mixture, where $\mathbf{x}_i \sim \frac{1}{2}\mathsf{N}(\vec{3}, I_p) + \frac{1}{2}N(-\vec{3}, I_p)$; this was shown in Figure 2b, and is analyzed in more detail in Figure 5.
3. speech data, a real dataset retrieved from OpenML, shown in 2d, analyzed in more detail in Figure 6;
4. unresidualized returns, where each row contains the unresidualized return over a 30 minute interval, shown in Figure 7;
5. speech data with Signal-PC alignment, shown in Figure 2e shown again in Figure 8;

As a refresher for notation:

- We study ridge regression given data $(\mathbf{X}, \mathbf{y})$, where $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$. The training set $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O} \in \mathbb{R}^{n\times p}$ is right-rotationally invariant, meaning $\mathbf{O} \sim \text{Haar}(\mathbb{O}_p)$. The error $\boldsymbol{\epsilon}$ is centered with independent entries.
- We are interested in the test loss on a test set $\widetilde{\mathbf{X}} = \widetilde{\mathbf{Q}}^\top \widetilde{\mathbf{D}}\widetilde{\mathbf{O}} \in \mathbb{R}^{n'\times p}$, defined as

$$R_{\mathbf{X},\mathbf{y}}(\hat{\boldsymbol{\beta}}(\mathbf{X},\mathbf{y}), \boldsymbol{\beta}) = \frac{1}{n'}\mathbb{E}\left[\|\widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\|^2 \mid \boldsymbol{X}, \boldsymbol{y}\right].$$

- The two key parameters are the signal strength $r^2 = \|\boldsymbol{\beta}\|_2^2/n$ and the noise level $\sigma^2 = \mathbb{E}[\boldsymbol{\epsilon}_i^2]$
- Let $\mathbf{o}_i^\top$ denote the i-th row of $\mathbf{O}$ and $\widetilde{\mathbf{o}}_i$ denote the i-th row of $\widetilde{\mathbf{O}}$.
- **Aligned eigenvectors** refer vectors o_i which align with the signal β. They are indexed by $\mathcal{J}_a$, and satisfy $\boldsymbol{\beta} = \sum_{i\in\mathcal{J}_a} \sqrt{n}\alpha_i \mathbf{o}_i + \boldsymbol{\beta}'$.
- **Coupled eigenvectors** refer to vectors $\mathbf{o}_i$ which align with $\widetilde{\mathbf{o}}_i$, which are eigenvectors of the training covariates. They are indexed by $\mathcal{J}_c$, and we assume for each $i \in \mathcal{J}_c$, that $\mathbf{o}_i = \widetilde{\mathbf{o}}_i$.
- Aligned and coupled eigenvectors allow us to model better model structures in real data that may not initially fall under the coverage of our method. We will refer to scenarios covered without the need for coupled/aligned eigenvectors as the **simpler setting**; in the main text, these are shown in Figure 1. For scenarios that require coupled/aligned eigenvectors, we refer to them as the **coupled/aligned** setting; these are covered in Figure 2 of the main text.

The most important assumptions, reproduced informally from the main text, are as follows:

1. On the regularity of singular value distributions:
   - In the simpler setting: we require **Assumption 2**, that the maximum singular value $\lambda_{\max}(D)$ is almost surely bounded.

- In the aligned/coupled setting, we further require **Assumption 11** that the expected maximum singular value of the test covariates, $\|\mathbb{E}[\widetilde{\mathbf{D}}^\top \widetilde{\mathbf{D}}]\|_{\mathsf{op}}$, is also almost surely bounded.

How to check this: One can generally verify that this assumption is satisfied by plotting a histogram of the singular values $(D_{ii})_{i=1}^p$ and $(\widetilde{D}_{ii})_{i=1}^p$. In the examples we discuss, we only do this for the training set since we do not expect any distribution shift. As examples, see Figures 4a, 5a etc.

2. On the relation between the eigenvectors of the training set $\{\mathbf{o}_i\}_{i=1}^p$ and test set $\{\widetilde{\mathbf{o}}_i\}_{i=1}^p$:
   - In the simpler setting, we have **Assumption 6**, which states the test set is independent of the training set.
   - In the coupled setting, we have **Assumption 7**, which states the top eigenvectors of the test and training set are exactly equal, i.e. for $i \in \mathcal{J}_c$, we have $o_i = \tilde{o}_i$.

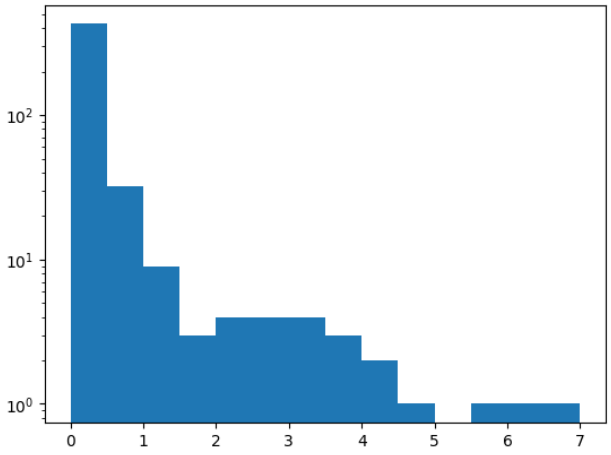
   One should think of Assumption 7 as a weakening of Assumption 6, where we allow for additional forms of structure to handle trends we see in data.

   How to check this: We compute the overlaps $\sqrt{p}\langle \tilde{o}_i, o_j \rangle$ for the top few eigenvectors (we choose 10). Note that if these two vectors were independent, then one has $\sqrt{p}\langle \tilde{o}_i, \tilde{o}_j \rangle \xrightarrow{p\to\infty} N(0,1)$. Hence one can empirically observe if two eigenvectors possess overlap that is significantly larger than expected; if they do, then they should be coupled. For examples, see Figures 4c, 5c.

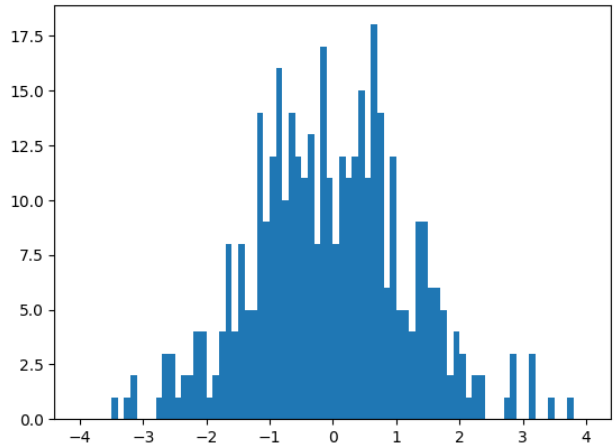
   One limitation of this approach is that each eigenvector of the training set can only be coupled to one in the test set; what this means is that if we find eigenvectors in the training set that overlap heavily with multiple eigenvectors of the test set, then our approach is not expected to perform well. We will see this occur in setting 4.

3. On the position of β relative to $\{o_i\}_{i=1}^p$, which are the rows of O:
   - In the simpler setting, we have **Assumption 5**, which states that $\{o_i\}_{i=1}^p$ are independent of β.
   - In the aligned setting, we have **Assumption 8**, which states a finite collection of eigenvectors indexed by $\mathcal{J}_a$ are aligned with β, so that $\beta = \sum_{i\in\mathcal{J}_a} \sqrt{n}\alpha_i o_i + \beta'$.

   How to check this: Identification of alignment is discussed in Li and Sur (2023), which proposes a hypothesis testing framework for identifying J_a. They provide a method to compute a p-value per eigenvector that tests whether or not it is aligned; one can then apply the Benjamini-Hochberg procedure to control the false discovery rate. In the 5th example, we use the same approach, where the p-values are shown in Table 3.



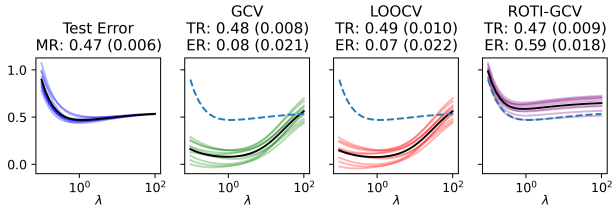
(a) Histogram of singular values $(D_{ii})_{i=1}^n$



(b) Histogram of overlaps $\{\sqrt{p}\langle\tilde{o}_i, o_j\rangle : 1 \le i \le j \le 20\}$

	$\tilde{o}_1$	$\tilde{o}_2$	$\tilde{o}_3$	$\tilde{o}_4$	$\tilde{o}_5$	$\tilde{o}_6$	$\tilde{o}_7$	$\tilde{o}_8$	$\tilde{o}_9$	$\tilde{o}_{10}$
o_1	2.87	2.02	1.04	2.18	1.48	1.37	2.34	0.58	1.64	2.30
o_2	0.31	1.03	0.15	3.20	0.78	0.52	0.82	0.35	0.43	0.76
o_3	2.66	3.17	0.07	1.42	3.47	0.44	0.48	0.98	0.73	1.01
o_4	0.20	0.74	2.73	1.07	0.85	1.82	0.46	0.25	0.76	1.64
o_5	1.32	0.12	1.99	1.66	1.12	1.75	0.54	0.40	0.32	1.81
o_6	2.37	0.57	1.70	0.38	1.26	1.07	0.63	0.45	0.39	1.72
o_7	0.93	1.86	0.23	0.84	0.48	0.02	0.69	0.02	0.63	1.41
o_8	4.12	0.81	2.57	1.32	1.52	1.03	0.88	1.48	0.72	0.88
o_9	0.14	1.01	0.01	1.17	1.27	2.43	1.14	0.45	0.32	1.12
o_{10}	0.02	1.28	1.64	0.09	0.94	0.50	0.42	1.32	0.34	0.66

(c) Numerical values of overlaps; Cell (i, j) contains the value of $\sqrt{p}|\langle o_i, \tilde{o}_j\rangle|$. Note that $\sqrt{p}\langle o_i, \tilde{o}_j\rangle \xrightarrow{p\to\infty} N(0, 1)$



(d) Tuning curves for Residualized Returns; same as in Figure 2 of the main text.

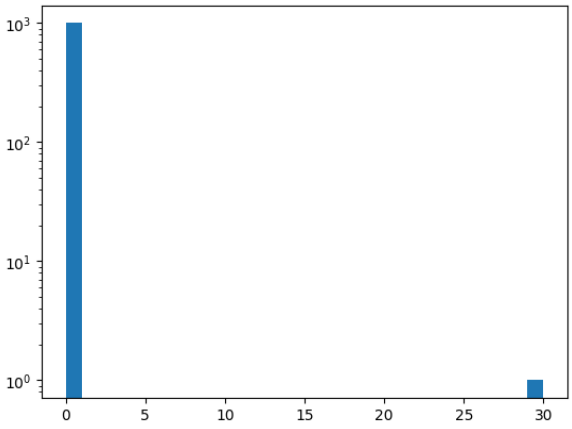
Figure 4: Residualized returns setting, $n = p = 493$, $r^2 = \sigma^2 = 1$; Figure 4b is essentially a histogram of values in 4c.

10 Residualized Returns

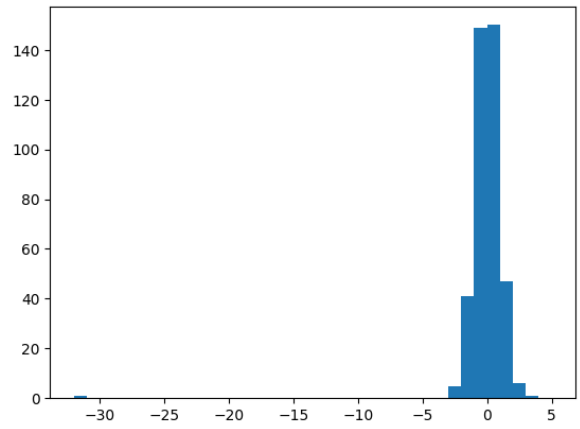
Overview of Setting: to recap from the main text (Figure 2d), we considered data where each row contained the 30 minute residualized returns of a collection of 493 stocks. Here, we set $n = p = 493$ and $r^2 = \sigma^2 = 1$.

Checking Assumptions:

- Regularity of Singular Value Distribution: In Figure 4a, we plot a histogram of the singular values and see that it is are well-behaved with no strong outliers.
- On coupling: Furthermore, while the histogram of singular values (shown in Figure 4b, numerical values in 4c) has larger tails than expected for a normal distribution, they are not overly so, and thus we find we do not require coupling either (i.e. Assumption 6 essentially holds). As a result, applying our method without coupling performs well (result shown in Figure 4d.



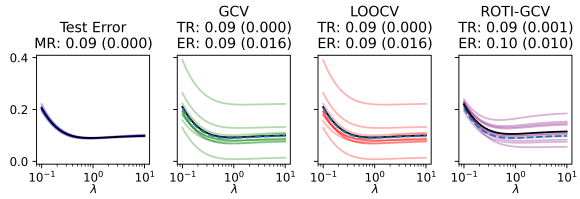
(a) Histogram of singular values $(D_{ii})_{i=1}^n$



(b) Histogram of overlaps $\{\sqrt{p}\langle\tilde{o}_i, o_j\rangle : 1 \le i \le j \le 20\}$. **Note the outlier value at around -31.**

	$\tilde{o}_1$	$\tilde{o}_2$	$\tilde{o}_3$	$\tilde{o}_4$	$\tilde{o}_5$	$\tilde{o}_6$	$\tilde{o}_7$	$\tilde{o}_8$	$\tilde{o}_9$	$\tilde{o}_{10}$
o_1	31.62	0.01	0.00	0.01	0.03	0.00	0.03	0.02	0.01	0.01
o_2	0.00	0.71	0.35	0.32	1.67	0.22	0.92	0.37	1.09	0.36
o_3	0.02	0.38	1.23	0.78	0.60	1.44	0.04	0.86	0.56	0.26
o_4	0.01	0.74	0.07	0.61	0.30	1.50	0.29	0.41	0.65	0.11
o_5	0.01	0.57	1.05	0.27	0.21	0.36	0.97	0.58	0.99	0.47
o_6	0.02	0.37	0.49	0.37	1.31	0.67	0.94	2.08	0.16	1.24
o_7	0.01	0.18	1.20	0.22	0.02	1.12	0.78	0.91	0.81	2.33
o_8	0.01	0.33	0.02	0.72	0.87	1.49	0.25	2.44	0.58	1.27
o_9	0.03	1.55	0.54	0.92	0.43	0.13	0.46	0.45	0.33	0.04
o_{10}	0.00	0.36	1.01	1.65	0.95	1.15	0.83	0.75	0.83	0.80

(c) Numerical values of overlaps; Cell (i, j) contains the value of $\sqrt{p}|\langle o_i, \tilde{o}_j\rangle|$. Note that $\sqrt{p}\langle o_i, \tilde{o}_j\rangle \xrightarrow{p\to\infty} N(0, 1)$



(d) Tuning curves for Gaussian mixtures

Figure 5: Gaussian mixture setting, $n = p = 1000$; Figure 5b is essentially a histogram of values in 5c.

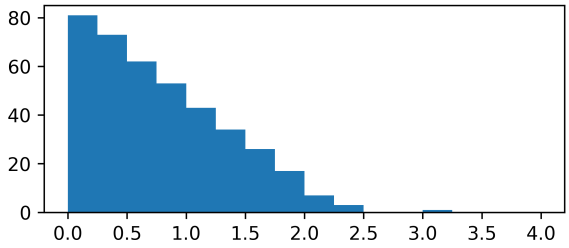
11 Gaussian mixture

To build intuition for the coupled eigenvectors condition (Assumption 7), we take a first look at how these diagnostic plots change when the data is instead drawn from the Gaussian mixture discussed in the main text (restated below):

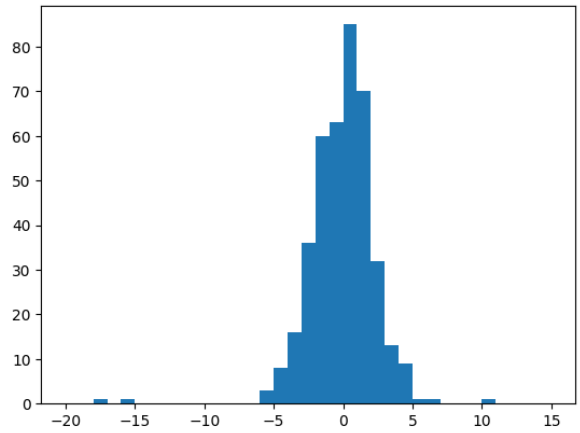
Overview of setting: Each row $x_i \sim \frac{1}{2}N(\vec{3}, I_p) + \frac{1}{2}N(-\vec{3}, I_p)$. We take $n = p = 1000$, and $r^2 = \sigma^2 = 1$.

Checking assumptions:

- Regularity of Singular Value Distribution: As seen in Figure 5a, we have one extremely large singular value, showing Assumption 2 is violated. This is usually a bad sign for our method, but in this case it manages to succeed, showing some robustness.
- On coupling: In Figures 5b and 5c, the huge outlier overlap is indication that the top two eigenvectors overlap with one another, meaning Assumption 6 does not hold. On the other hand, all the other eigenvectors do not align with each other. This is exactly the situation suited for our coupled eigenvectors condition (Assumption 7). As a result, coupling o_1 with $\tilde{o}_1$ produces solid results, shown in Figure 5d.



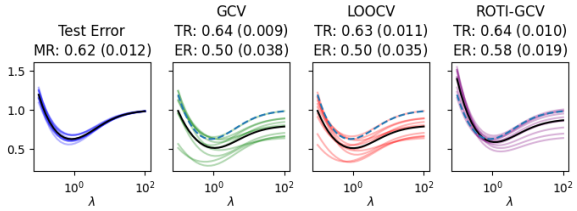
(a) Histogram of singular values $(D_{ii})_{i=1}^{n}$



(b) Histogram of overlaps $\{\sqrt{p}\langle\tilde{o}_i, o_j\rangle : 1 \le i \le j \le 20\}$

	$\tilde{o}_1$	$\tilde{o}_2$	$\tilde{o}_3$	$\tilde{o}_4$	$\tilde{o}_5$	$\tilde{o}_6$	$\tilde{o}_7$	$\tilde{o}_8$	$\tilde{o}_9$	$\tilde{o}_{10}$
o_1	17.30	2.72	0.63	2.01	1.55	1.02	0.30	0.96	1.70	0.27
o_2	0.83	15.24	0.68	1.63	1.34	1.22	0.89	1.42	0.27	2.19
o_3	1.08	1.93	10.93	1.60	0.64	2.02	4.63	0.35	4.14	2.75
o_4	0.29	2.78	2.92	0.22	3.98	6.63	2.98	1.65	0.54	1.56
o_5	1.27	2.70	4.02	2.14	5.11	1.29	4.34	4.72	2.95	2.24
o_6	1.19	0.56	4.82	5.95	0.37	3.83	1.37	0.87	4.98	1.04
o_7	2.21	0.70	2.14	2.65	4.84	1.77	0.54	1.00	1.65	2.53
o_8	0.46	0.15	2.96	0.16	1.66	4.49	0.22	0.40	1.82	0.10
o_9	0.30	0.40	0.72	0.01	1.63	1.39	3.86	4.53	1.12	2.57
o_{10}	1.14	2.51	0.90	0.77	0.37	0.79	2.44	1.83	1.81	2.34

(c) Numerical values of overlaps; Cell (i, j) contains the value of $\sqrt{p}|\langle o_i, \tilde{o}_j\rangle|$. Note that $\sqrt{p}\langle o_i, \tilde{o}_j\rangle \xrightarrow{p\to\infty} N(0, 1)$



(d) Tuning curves for Speech Data

Figure 6: Speech data setting, $n = p = 400$, $r^2 = \sigma^2 = 1$; Figure 6b is essentially a histogram of values in 6c.

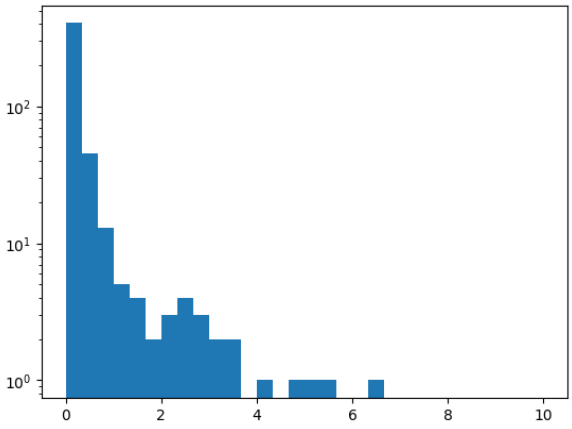
12 Speech Data

We find a good use case for our coupled eigenvector condition on speech data taken from OpenML. This is a real dataset which is not included in the main manuscript.

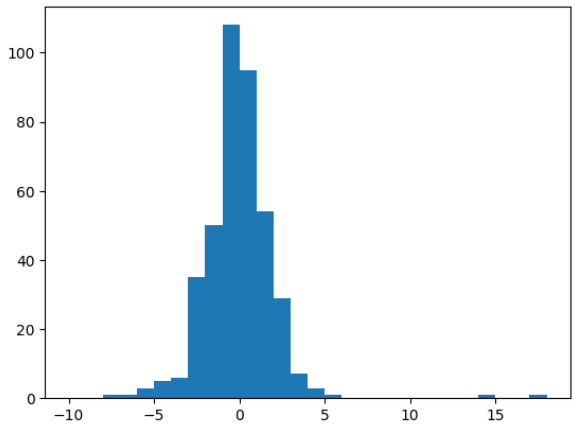
Overview of setting: Dataset sourced is speech data from Le (2017). $n = p = 400$, and $r^2 = \sigma^2 = 1$.

Checking assumptions:

- Regularity of Singular Value Distribution: Here, in Figure 6a, the spectrum has no outlier values.
- On coupling: Next, we observe that for each of the top 3 eigenvectors training set, they align strongly with the corresponding eigenvector of the test set, and none of the others. This is exactly the setting in which our coupled eigenvectors can be used. Furthermore, each of the top train eigenvectors is strongly aligned with only one of the test eigenvectors. In light of Table 6c, we choose to couple the first 3 eigenvectors with the respective test-set eigenvectors. This produces the result shown in Figure 6d, showing that it successfully captures the required structure. This result is quite robust to the choice of J_c; using 6 eigenvectors or even 10 does not change the outcome much.



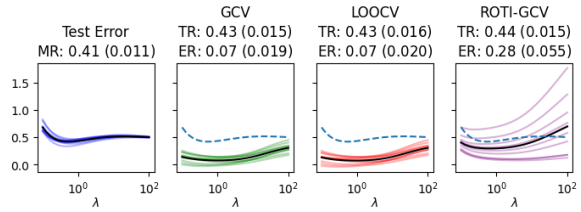
(a) Histogram of singular values $(D_{ii})_{i=1}^n$



(b) Histogram of overlaps $\{\sqrt{p}\langle\tilde{o}_i, o_j\rangle : 1 \le i \le j \le 20\}$

	$\tilde{o}_1$	$\tilde{o}_2$	$\tilde{o}_3$	$\tilde{o}_4$	$\tilde{o}_5$	$\tilde{o}_6$	$\tilde{o}_7$	$\tilde{o}_8$	$\tilde{o}_9$	$\tilde{o}_{10}$
o_1	17.23	2.43	2.84	0.98	2.93	0.55	0.27	0.84	4.29	2.83
o_2	0.72	14.25	1.22	0.23	5.62	1.36	2.26	2.36	0.49	0.50
o_3	6.28	2.60	4.16	2.43	0.95	3.18	0.77	4.40	3.59	1.18
o_4	0.83	2.23	0.66	1.02	0.20	1.12	1.39	2.90	2.41	0.37
o_5	5.87	7.61	4.22	0.59	4.00	0.45	0.10	1.23	0.19	0.54
o_6	0.81	1.30	0.53	2.43	2.27	0.47	5.56	1.78	3.29	0.72
o_7	0.09	0.22	2.07	2.76	0.38	1.69	4.25	2.30	1.87	1.14
o_8	1.50	0.44	0.91	3.61	1.33	4.63	0.01	0.32	0.29	0.43
o_9	2.62	1.31	0.14	2.42	1.66	2.42	2.62	0.08	0.21	0.38
o_{10}	0.92	2.82	1.72	2.53	0.99	0.16	0.70	1.14	0.46	0.67

(c) Numerical values of overlaps; Cell (i, j) contains the value of $\sqrt{p}|\langle o_i, \tilde{o}_j\rangle|$. Note that $\sqrt{p}\langle o_i, \tilde{o}_j\rangle \xrightarrow{p\to\infty} N(0, 1)$



(d) Tuning curves for unresidualized returns

Figure 7: Poorly behaved setting, unresidualized returns, $n = p = 493$; Figure 7b is essentially a histogram of values in 7c

13 Poorly behaved case

We will now look at a case that behaves poorly for our estimator. Here, we look at unresidualized returns.

Overview of Setting: we consider data where each row contains 30 minute returns, but without residualization.

Checking Assumptions:

- Regularity of singular values: Here, in Figure 7a, there are some larger singular values, but they are not extreme outliers.
- On coupling: We see that when checking the alignment of test and training eigenvectors in Figures 7b and 7c, the top eigenvector of the test set aligns heavily with mulitple eigenvectors of the training set. This is a scenario in which we cannot use the coupled eigenvector condition, since we can only couple one train eigenvector with one test eigenvector. As a result, we see in Figure 7d that our method does not do well in this scenario, which is expected. Note that this example illustrates that these diagnostics can be used ahead of time to understand whether or not our method should succeed.

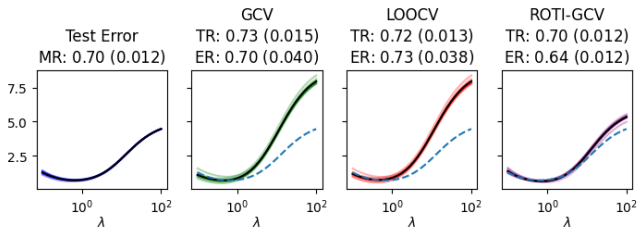


Figure 8: Speech data with both alignment and coupling

Table 3: BHq adjusted p-value for alignment.

	p
o_1	0.000
o_2	0.000
o_3	0.000
o_4	0.000
o_5	0.000
o_6	0.934
o_7	0.588
o_8	0.651
o_9	0.913
o_{10}	0.395

14 Speech data with both alignment and coupling

As a final example that includes both alignment and coupling, we perform another experiment with speech data that additionally contains Signal-PC alignment.

Overview of setting: The dataset is again the speech data from Le (2017). Recall that an aligned signal takes the form $\beta = \sum_{i \in J_c}^{n} \sqrt{n}\alpha_i + \beta'$, where the α_i are the weights of the aligned components, and β' is the unaligned component of the signal. We take $\alpha_i = 1/2$ for $i \in J_a = \{1, 2, 3, 4, 5\}$.

Checking assumptions:

- Singular value histogram and coupling detection is equivalent to that of Section 12.
- Alignment: Here we next identify alignment through the hypothesis testing framework of Li and Sur (2023). We include in Table 2 the Benjamini-Hochberg adjusted p-values of alignment of each component, finding that p-values for each of $i = 1, 2, 3, 4, 5$ are extremely small ($< 10^{-2}$), and those of the unaligned components are large (> 0.35), meaning we perfectly identify the aligned portion. Our method then performs well in this setting, as shown in Figure 8.