Dimensionality reduction, regularization, and generalization in overparameterized regressions

Ningyuan (Teresa) Huang*, David W. Hogg†, and Soledad Villar*

Abstract. Overparameterization in deep learning is powerful: Very large models fit the training data perfectly and yet often generalize well. This realization brought back the study of linear models for regression, including ordinary least squares (OLS), which, like deep learning, shows a "double-descent" behavior: (1) The risk (expected out-of-sample prediction error) can grow arbitrarily when the number of parameters p approaches the number of samples n, and (2) the risk decreases with p for $p > n$, sometimes achieving a lower value than the lowest risk for $p < n$. The divergence of the risk for OLS can be avoided with regularization. In this work, we show that for some data models it can also be avoided with a PCA-based dimensionality reduction (PCA-OLS, also known as principal component regression). We provide non-asymptotic bounds for the risk of PCA-OLS by considering the alignments of the population and empirical principal components. We show that dimensionality reduction improves robustness while OLS is arbitrarily susceptible to adversarial attacks, particularly in the overparameterized regime. We compare PCA-OLS theoretically and empirically with a wide range of projection-based methods, including random projections, partial least squares (PLS), and certain classes of linear two-layer neural networks. These comparisons are made for different data generation models to assess the sensitivity to signal-to-noise and the alignment of regression coefficients with the features. We find that methods in which the projection depends on the training data can outperform methods where the projections are chosen independently of the training data, even those with oracle knowledge of population quantities, another seemingly paradoxical phenomenon that has been identified previously. This suggests that overparameterization may not be necessary for good generalization.

1. Overparameterization and robustness in regression.

One of the most remarkable properties of contemporary machine-learning methods—and especially deep learning—is that models with enormous capacity nonetheless generalize well. Overparameterized models have the flexibility to perfectly fit any training data, but (in many cases) still make good, non-trivial predictions on held-out or test data. Those good predictions contradict both our folklore and our intuitions.

The realization that overparameterization is good for machine learning led to a reconsideration of classical linear regressions. It turns out that even linear regressions can generalize well in the overparameterized regime; that is, when the number of parameters p far exceeds the number of training data points n (provided that the fitting is performed in a min-norm or regularized way that limits the coefficients in the unconstrained $(p - n)$-dimensional subspace). Both linear regressions and more complex machine-learning methods typically show a "double-descent" phenomenon, recently identified by Belkin et al. [4]: (1) The underparameterized and overparameterized regimes are separated by a "peaking" phenomenon [22], or "jamming peak" [16, 52], in which the "risk"—the expected out-of-sample prediction error—blows up when the model capacity just reaches overfitting (at $p = n$ in the linear case); (2) The risk further decreases with the number of parameters in the overparameterized regime, sometimes achieving a lower value than the underparameterized regime.

*Department of Applied Mathematics and Statistics, Johns Hopkins University, and Mathematical Institute for Data Science, Johns Hopkins University.

†Flatiron Institute, a division of the Simons Foundation, and Center for Cosmology and Particle Physics, Department of Physics, New York University.

Correspondence should be addressed to soledad.villar@jhu.edu

The double-descent behavior raises the following research questions: Can we avoid the peaking phenomenon (RQ1)? Is overparameterization necessary for good generalization (RQ2)? The understanding of this phenomenon in the context of linear regression has provided some perspective on deep learning. An important result common to both deep learning and linear regression shows that regularization is important in or near the overparameterized regime [17, 13, 29, 60, 36, 43, 39, 9, 57, 20].

To address RQ1, we study double-descent in the context of linear regression, where the peaking phenomenon has a simple explanation that involves the (equivalent of the) condition number of the training features (the ratio of the largest singular value to the smallest). Prior work shows that the peaking phenomenon disappears with regularization using ridge penalty [17]. In this work, we show that it also disappears with dimensionality reduction, another canonical form of regularization.

In some sense—and in the attitude we will take here—the peaking phenomenon at $p \approx n$ is a kind of lack-of-robustness. Similarly, susceptibility to adversarial attacks is also a kind of lack-of-robustness. These things ought to be related: In our view, a robust regression will not have divergent risk nor be extremely susceptible to attack. We connect these ideas here, and note that some models that generalize well nonetheless are extremely weak against adversarial attacks. In the overparameterized regime the default linear model (ordinary least squares) makes good predictions but is not robust in the sense that it is arbitrarily susceptible to attack.

Regressions have been attacked adversarially in a few ways. We focus here on attacks against the training features, but there are also attacks against the training labels [42], and attacks against the test data at test time. We limit our discussion on the data-poisoning attack—that is, adding one adversarial data point to the training data [8, 33]. We refer the readers to the recent works in [32, 33, 24] for other forms of attacks including the Rank-1 (and Rank-k) attacks or the adversarial perturbation attacks.

There are different ways to measure success for attacks against regressions, including increases in the risk [24], distortion of the regression coefficients [33], and other kinds of distortions to the properties of the data, e.g., [46]. Here we are focused on robustness and prediction, so we care most about the risk. Connected to our motivation and results, there is also adversarial training, which has been developed as a kind of regularization for regressions; it protects regressions from attack and also makes the peaking phenomenon disappear [24]. On the other hand, deep generative classifiers—that produce a generative model for the training data, similar to a low dimensional parameterization of the data distribution—are shown to be more robust against adversarial attacks than deep discriminative classifiers [35, 25]. We design a simple generative model for linear regression and demonstrate how it could act as an implicit regularization and avoids the peaking phenomenon.

To investigate RQ2, we adopt the framework of projection-based methods, where the input data can be projected to a higher-dimensional feature space (i.e., overparameterization), or a lower-dimensional one (i.e., dimensionality reduction). This framework can be recast as a two-layer (linear) neural networks [2], where the first layer performs the projection (not trained), and the second layer performs linear regression (trained). We compare the risk behavior of multiple projection-based methods, including ordinary least squares after a PCA-based dimensionality reduction (PCA-OLS; also sometimes known as principal component regression), partial least squares, random projections, and classes of generative models and latent-variable models. All projection methods we consider herein involve transforming the input $X \in \mathbb{R}^{n \times p}$ linearly to some feature embeddings in $\mathbb{R}^{n \times k}$, followed by a linear regression on the transformed

features. Since the regression takes the transformed features, the interpolation threshold (i.e., the peaking) appears at $k \approx n$ (instead of $p \approx n$). In this setting, overparameterization (i.e., $k > \min\{n, p\}$) is only possible when the projection matrix is independent of the training data (recall PCA-OLS is only possible for $k \leq \min\{n, p\}$). Previous work has shown that the (individual) risk of data-independent projection methods monotonically decreases with k when $k > n$ [62, 2, 59]. However, it is not clear whether these overparameterized projection methods outperform their classic counterparts that choose the projection based on the training data, such as PCA-OLS and partial least squares.

We summarize our motivations and our contributions in Section 3, after we give some problem setup and define some forms for linear regression in Section 2. We follow that with analytical results in dimensionality reduction in Section 4, and a comparison with other projection-based regression models in Section 5. In Section 6 we discuss analytical results for adversarial attacks in the context of robustness of OLS in comparison with PCA-OLS, and in Section 7 we provide numerical experiments.

2. Linear regression: Problem setup and methods. Let $\{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. We imagine having n data points (x, y) that (unknown to us) were generated from a joint Gaussian $\mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_x, \mu_y) = (0_p, 0_1)$ and $\Sigma = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix}$. In other words $p(x, y) = \mathcal{N}(\mu, \Sigma)$, and therefore

$$\mathbb{E}_y[y \,|\, x] = C_{yx}\, C_{xx}^{-1}(x - \mu_x) + \mu_y \tag{2.1}$$

(see for instance [48] appendix A). In the case where $\mu = 0$, this generative model is equivalent to $x \sim \mathcal{N}(0, C_{xx})$ and $y = x^\top \beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and

$$\beta := C_{xx}^{-1}\, C_{xy}, \quad \sigma^2 := C_{yy} - C_{yx}\, C_{xx}^{-1}\, C_{xy}\,. \tag{2.2}$$

This is now the standard linear generative model from the literature, with standard parameters β and σ. Let $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times 1}$ be training data in rectangular form.

We can consider several regression methods for finding an estimate $\hat{\beta}$ for the linear parameters β. Different estimators are compared in terms of their risk, that for our purposes will be defined as the expected squared error for out-of-sample predictions:

$$\mathcal{R}(\hat{\beta}) = \mathbb{E}_{Y, x_*, y_*}[\|x_*^\top \hat{\beta} - y_*\|_2^2 \,|\, X] \tag{2.3}$$

$$= \mathbb{E}_{Y, x_*}[\|x_*^\top (\hat{\beta} - \beta)\|_2^2 \,|\, X] + \sigma^2\,, \tag{2.4}$$

where (x_*, y_*) are test points, fresh samples from the same distribution.

With this problem setup, there are many possible methods for performing linear regression:

Ordinary least squares (OLS): This finds the linear combination of features X that best predict the labels Y in a least-square sense: $\hat{\beta}_{\text{OLS}} = \arg\min_\beta \|X\beta - Y\|_2^2$. In the overparameterized regime it chooses from among equivalent alternatives the min-norm solution. We obtain $\hat{\beta}_{\text{OLS}}$ by computing $X^\dagger Y$, where $X^\dagger$ denotes the pseudo-inverse of X (the inverse that inverts only the non-zero eigenvalues of the matrix), namely:

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y \quad \text{if } p < n \tag{2.5}$$

$$\hat{\beta}_{\text{OLS}} = X^\top (X\, X^\top)^{-1} Y \quad \text{if } p > n\,, \tag{2.6}$$

assuming that $\text{rank}(X) = \min\{p, n\}$.

Ridge regression: This is a variant of least-squares, but with an l_2 penalty on the regression coefficients, regularizing the fit: $\hat{\beta}_\lambda = \arg\min_\beta \|X\beta - Y\|_2^2 + n\lambda\|\beta\|_2^2$. There are other kinds of penalties by constraining the norm of the estimator, for instance l_1 (the Lasso), elastic net.

PCA-OLS: In this case we perform OLS, but—before starting—reduce the rank of the training data by performing a PCA-based dimensionality reduction:

$$\hat{\beta}_{\mathrm{PCA},k} = \arg\min_\beta \|X_{\mathrm{PCA},k}\,\beta - Y\|_2^2\,, \tag{2.7}$$

where $X_{\mathrm{PCA},k}$ is the rank-k PCA approximation to X, with $k < \min\{n, p\}$. There are other equivalent formulations to PCA-OLS like the one in [14]. In our formulation,

$$\hat{\beta}_{\mathrm{PCA},k} = (X_{\mathrm{PCA},k}^\top X_{\mathrm{PCA},k})^\dagger\, X_{\mathrm{PCA},k}^\top\, Y. \tag{2.8}$$

We remark that PCA-OLS is also known as Principal Component Regression (PCR) in the literature. In [62], Xu and Hsu studied the case where the true population covariance is known and the projection is onto the k principal components of the population covariance. We shall call this regression method *oracle*-PCR.

Partial least squares (PLS): This is a dimensionality-reduction based regression similar to PCA-OLS. PLS not only maximizes the variance of the projected features as PCA-OLS, but also the covariance of the projected responses and projected features. PLS is widely studied in the chemometrics and statistics literature [58, 18, 12]. The general form of PLS can be written as:

$$X \approx TP^\top;\ T \in \mathbb{R}^{n\times k}, P \in \mathbb{R}^{p\times k}. \tag{2.9a}$$

$$Y \approx UQ^\top;\ U \in \mathbb{R}^{n\times k}, Q \in \mathbb{R}^{q\times k}. \tag{2.9b}$$

When Y is an univariate response variable, as in our analysis, PLS can be formulated as projecting the features to a Krylov space, followed by OLS [49]:

$$\hat{\beta}_{\mathrm{PLS,k}} = \arg\min_\beta \|\,(X\,\Pi_{X_{\mathrm{PLS}}})\,\beta - Y\|_2^2\,, \tag{2.10}$$

where $\Pi_{X_{\mathrm{PLS}}} = [s_{xy}, As_{xy}, A^2 s_{xy}, \cdots, A^{k-1}s_{xy}]$, $s_{xy} := X^\top Y, A := X^\top X$, and $[\cdot]$ denotes column concatenation. Note that OLS, ridge regression, PCA-OLS and PLS can be unified under the framework of continuum regression [53, 28].

Random projection methods: PCA and PLS project the original features X via a *data-dependent* projection matrix $\Pi \in \mathbb{R}^{p\times k}$ that is constructed from the training data. Other classes of methods of interest use random projections Π, chosen independently of the training data. All these projection methods can be written as

$$\hat{\beta}_\Pi = (X\Pi)^\dagger Y. \tag{2.11}$$

For example, the random orthogonal projection in [37] samples Π uniformly over the set of orthogonal matrices such that $\Pi^\top\Pi = I_k$ for $k \le p$ (or $\Pi\,\Pi^\top = I_p$ for $k \ge p$); the random Gaussian projection in [2] chooses $\Pi = [w_1, \cdots, w_k]$, where $w_i \overset{i.i.d}{\sim} \mathcal{N}(0, p^{-1}I_p)$.

Data-dependent projections enforce the rank of Π be at most n, while random projection allows $\mathrm{rank}(\Pi) = \min\{p, k\}$ to exceed n if $p > n, k > n$. In this case, instead of reducing feature dimensions, random projection lifts the original features to a higher-dimensional space, which

is key to kernel-based learning and deep neural networks. For example, [2] identifies this model as a linear two-layer neural network where the first layer is random (untrained) and it performs a random Gaussian projection and the network is optimized only through the second layer parameters.

Generative and latent-variable models: In the classical machine learning literature, *generative* approaches are those that attempt to learn the joint distribution of the observable variable X and the target variable Y [45], or the distribution of X conditioned on Y. We hint towards this approach earlier in Section 2, where we construct the data from a joint distribution $p(x, y)$ and we show how the generative formulation translates to the more common—discriminative—linear regression setting $Y = X\beta + \epsilon$.

The generative regression model we propose—at training time—generates the features X and the targets Y as a linear function of latent variables $Z \in \mathbb{R}^{n\times k}$. In particular we aim to find Z, P, Q such that $Y \approx ZP$, and $X \approx ZQ$, where Z are the latent variables, and $P \in \mathbb{R}^{k\times 1}$, $Q \in \mathbb{R}^{k\times p}$ are linear operators.

Note that if (Z, P, Q) is a generative model for (X, Y), so is $(ZS, S^{-1}P, S^{-1}Q)$ for any S invertible $k \times k$ matrix. Therefore we set P to be a $k \times 1$ projection matrix given by the user, and we train the model to find Z and Q.

We define the generative linear regression as follows: $\hat{\beta}_{generative} = \hat{Q}^{\top\dagger}P$ where P is a $k\times 1$ projection matrix given by the user, Q is a $k \times p$ operator, and Z is a $n \times k$ matrix of latent variables. Estimates for Q and Z are found by

$$\hat{Q}, \hat{Z} = \underset{Q,Z}{\operatorname{argmin}} \|X - Z\,Q\|_2^2 \text{ subject to } Z\,P = Y. \tag{2.12}$$

We train the model by solving (2.12) via alternately optimizing with respect to Q and Z.

Our generative model can be viewed as fitting PLS when choosing the same projection matrix for both X and Y (i.e., $T = U \equiv Z$ in equation (2.9)). This is similar to the latent space model in [17], where they relax the constraint to be $ZP \approx Y$. They provide an asymptotic risk analysis by simplifying Q as an orthogonal projector and assuming the latent variables Z from isotropic Gaussian distribution. Our generative model also reduces to PCA-OLS if the constraints are removed.

3. Condition numbers, risk, and susceptibility to attack: Our contributions. Because $\hat{\beta}_{OLS}$ involves an inverse (or matrix solve or pseudo-inverse) of the $X^\top X$ or $X\,X^\top$ (which are related closely to the empirical variance of the features), the risk—the expected out-of-sample squared prediction error—will be strongly dependent on the condition number of the empirical variance. In general, the risk will get large as the condition number gets large. And indeed, the peaking phenomenon is related to the expectation of the condition number of this empirical variance [17].

Ridge regression has been shown to avoid the peaking phenomenon [17], and it does so by adding $n\,\lambda$ to the diagonal of the $X^\top X$ or $X\,X^\top$ matrix in the $\hat{\beta}_\lambda$ expression. This limits the condition number, makes the inverse well behaved, and limits the risk.

We conjecture that any regression method that controls or limits the condition number of the empirical covariance of the features will avoid or ameliorate the peaking phenomenon. This leads us to consider the PCA-OLS method, which replaces X with a dimensionality-reduced copy of X, which thereby formally has infinite condition number, but in the context of a pseudo-inverse has a well-behaved effective condition number, so long as the k-th largest eigenvalue of empirical covariance matrix is well bounded away from 0. (The effective condition number here is the ratio of the largest eigenvalue of the matrix to the smallest *non-zero*

eigenvalue.) We conjecture that generative-model regressions (described above) will avoid the peaking for the same reasons. This also motivates us to analyze PCA-OLS under different data generating process and compare it with other projection-based methods, some of which do not control the condition number.

In what follows (Section 4), we provide matching upper and lower bounds on the risk for PCA-OLS, in the setting where the "effective rank" $r_0(C_{xx}) := \frac{\mathrm{tr}(C_{xx})}{\lambda_1} = o(n)$. The notion of effective rank is particularly useful in the analysis of overparameterized models in linear regression [3] and principal component analysis [31]. It is also closely related to the study of basis expansion methods, which we further discuss in Section 5.3. Under the setting $r_0(C_{xx}) = o(n)$, we show that PCA-OLS remains bounded for all $k < n$ as long as the k-th largest population eigenvalue is bounded sufficiently away from 0, whereas unregularized OLS further requires the population covariance C_{xx} to have a heavy tail [3], otherwise the risk of OLS blows up at $n \approx p$ [17]. This answers RQ1: Dimensionality reduction as a form of regularization can avoid the peaking phenomenon. We demonstrate our results in Section 7.1.

In Section 5, we consider various projection-based methods and discuss their theoretical properties. Our analysis is supported by extensive experiments in Section 7.2. Using our framework introduced in Section 1, we vary the choice of projection dimension k to compare the risk behaviors of these methods, where overparameterization occurs when $k > n$ (given $p > n$). Remarkably, we empirically observe that projection methods independent of the training data, like oracle-PCR and random projections that can overparameterize, perform worse than projection methods based on the data, such as PCA-OLS. Although data-independent projection methods show decreasing risk with further overparameterization, in practice, they must be coupled with regularization to generalize well [63, 37] (and they do generalize well when regularized, see Section 7.2). This answers RQ2: overparameterization is not necessary for good generalization. For example, PCA-OLS always perform dimensionality reduction where the optimal choice of $k < \min\{n, p\}$, and it seems to outperform all (unregularized) overparameterized methods.

Superficially, our result is in contrast with that of by Xu and Hsu [62], who perform a principal component regression and observe a double-descent behavior. But in fact there is no contradiction: That prior result is based not on an empirical PCA of the features; it is based on an oracle version of principal component regression (oracle-PCR), a method that (unrealistically) requires knowledge of the true (unobservable) generating distribution of the features, that is widely studied in the statistics literature [41, 15] and amenable to exact analysis using the Marchenko-Pastur distribution. Although oracle-PCR can sometimes yield smaller risk at $k > n$ (e.g. for isotropic covariance model), we empirically observe that it is no better than min-norm OLS (which is PCA-OLS when $k \geq \min\{n, p\}$, see Figure 7.4). Moreover, with high probability, oracle-PCR suffers from the peaking phenomenon at $k \to n$ due to unbounded variance [62, 59], where PCA-OLS has a bounded variance given that the k-th largest population eigenvalue λ_k is bounded away from 0. The fact that PCA-based estimates are more robust than ones derived with oracle-PCR (at least in the regime $n \approx k$) seems to be an instance where using the predictor of the covariance decreases the variance of estimators with respect to using the true value. This sort of paradoxical phenomenon has been identified in different contexts within the statistics literature [19, 55].

Going beyond benign training data, we consider "data-poisoning" attacks in Section 6, in which the attacker is permitted to add one data point (x_0, y_0) to the training data prior to training. We find that if that data point is carefully chosen to increase substantially the condition number of the empirical covariance of the features (by, say, introducing a small

but non-zero singular value), it also has a significant effect on the risk. We show that, in the OLS case, in the overparameterized regime $p > n$, the attack can be arbitrarily harmful to the risk, because it can arbitrarily increase the condition number of the empirical variance. This implies that overparameterized projection methods are also susceptible to attacks, unless they are properly regularized. We show that PCA-OLS and ridge regression are not nearly as susceptible to data-poisoning attacks, as expected, since they control the condition number of the matrix being inverted (or effective condition number of the matrix being pseudo-inverted). We conjecture that other kinds of regularized regressions, and the generative model defined in the previous section, will also be (at least partially) protected against such attacks. We show empirical evidence of such claims in Section 7.1.

4. Generalization properties of PCA-OLS. In this Section we analyze the risk of the estimator $\hat{\beta}_{\mathrm{PCA},k}$ and we provide an upper bound: The risk is independent of the number of parameters p and monotonically decreases with the number of samples n while number of principal components k is fixed.

We define the expected value of an estimator $\hat{\beta}$ of the form of (2.5) or (2.6) conditioned on the training data as

$$\tilde{\beta} := \mathbb{E}_Y[\hat{\beta} \,|\, X] = \Pi_X \, C_{xx}^{-1} \, C_{xy} = \Pi_X \, \beta \,, \tag{4.1}$$

where Π_X is the $p \times p$ orthogonal projection onto the span of X. We note that when Π_X is the identity or when the span of X contains the span of β, the estimator $\hat{\beta}$ is unbiased. Let $\Pi_{X_\perp}$ be the projection to the space orthogonal to the span of X.

Following the notation from [17], the risk for the OLS case (where $\hat{\beta} = \hat{\beta}_{\mathrm{OLS}}$) can be decomposed as a quadratic sum of bias plus variance in the following way:

$$\mathcal{R}(\hat{\beta} \,|\, X) = \underbrace{\mathbb{E}_{x_*}[(x_*^\top(\beta - \tilde{\beta}))^2 \,|\, X]}_{\text{bias squared}} + \underbrace{\mathbb{E}_{Y,x_*}[(x_*^\top(\hat{\beta} - \tilde{\beta}))^2 \,|\, X]}_{\text{variance}} + \sigma^2 \tag{4.2}$$

$$= \underbrace{\beta^\top \, \Pi_{X_\perp} \, C_{xx} \, \Pi_{X_\perp} \, \beta}_{\text{bias squared}} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left((\frac{1}{n} X^\top X)^\dagger \, C_{xx} \right)}_{\text{variance}} + \sigma^2, \tag{4.3}$$

Hastie et. al. [17] give an expectation for the risk as a function of p/n (in the limit of $n \to \infty$) making use of the Marchenko–Pastur distribution for eigenvalues of a random matrix. The key step of their argument writes $X = Z \, C_{xx}^{1/2}$, where Z is a standard spherical Gaussian. With this substitution and the cyclical property of the trace, the variance term becomes independent of C_{xx} and it reduces to the integral of the inverse of the singular values of Z with respect to the Marchenko–Pastur measure. Unfortunately the transformation $X = ZC_{xx}^{1/2}$ does not interact well with the PCA projection, which prevents us to use the cyclic property of the trace to obtain a closed-form expression for the risk. However, we are able to provide non-asymptotic bounds.

In order to analyze the risk of the PCA-OLS estimator $\hat{\beta} = \hat{\beta}_{\mathrm{PCA},k}$ we write

$$\tilde{\beta} = \Pi_{X_{\mathrm{PCA}}} \, \beta, \tag{4.4}$$

$$\mathcal{R}(\hat{\beta} \,|\, X) = \underbrace{\mathbb{E}_{x_*}[(x_*^\top(\beta - \tilde{\beta}))^2 \,|\, X]}_{\text{bias squared}} + \underbrace{\mathbb{E}_{Y,x_*}[(x_*^\top(\hat{\beta} - \tilde{\beta}))^2 \,|\, X]}_{\text{variance}} + \sigma^2 \tag{4.5}$$

$$= \underbrace{\beta^\top \, \Pi_{X_{\mathrm{PCA}\perp}} \, C_{xx} \, \Pi_{X_{\mathrm{PCA}\perp}} \, \beta}_{\text{bias squared}} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger \, C_{xx} \right)}_{\text{variance}} + \sigma^2, \tag{4.6}$$

where X_{PCA}, $\Pi_{X_{\text{PCA}}}$, and $\Pi_{X_{\text{PCA}\perp}}$ are the equivalents of their non-PCA counterparts for the rank-k PCA approximation to X. The proof of this statement is straightforward and we include it in Supplementary A.

In Theorem 1 we provide a coarse upper bound for the risk of PCA-OLS. The proof of this bound uses generic random matrix tools and makes no assumption on β nor the covariance matrix C_{xx}. In Theorem 2 we provide more refined upper and lower bounds for the risk that relies on the alignment of the top eigenspaces of empirical and population covariance matrices. Theorem 2 assumes the population covariance matrix C_{xx} and empirical covariance $\frac{1}{n}X^\top X$ satisfy a certain set of spectral gap assumptions from [31] and [40]. These assumptions hold when the number of samples is large enough, and the gap between distinct eigenvalues of the population covariance is not too small. We give a complete discussion later on this Section.

Theorem 1. *Let* $x_i \sim \mathcal{N}(0_p, C_{xx})$ $i = 1, \ldots, n$, *and*

$$y_i = x_i^\top \beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. *Let* c, t *be some constants,* λ_1 *be the largest eigenvalue of* C_{xx}, *and* $r_0(C_{xx}) := \frac{\operatorname{tr}(C_{xx})}{\lambda_1}$ *be the effective rank. Let* $M = c\lambda_1 \max\left\{\sqrt{\frac{r_0(C_{xx})}{n}}, \frac{r_0(C_{xx})}{n}, \frac{t}{n}\right\}$, *and assume* $M < \lambda_k$. *Then with probability greater than* $1 - e^{-t}$ *we have*

$$\mathcal{R}(\hat{\beta}_{PCA\text{-}OLS\text{-}k} \mid X) = \mathbb{B} + \mathbb{V} + \sigma^2, \tag{4.7}$$

$$\lambda_p \|\Pi_{X_{PCA\perp}}\beta\|^2 \leq \mathbb{B} \leq \|\beta\|^2\Big(M + \lambda_{k+1}\Big), \tag{4.8}$$

$$\frac{\sigma^2}{n}\frac{k\lambda_p}{\lambda_1 + M} \leq \mathbb{V} \leq \frac{\sigma^2}{n}\frac{k\lambda_1}{\lambda_k - M}, \tag{4.9}$$

where $\|\cdot\|$ *is the* 2*-norm for vectors, and* k *denotes the rank-*k *PCA with* $k < \min\{n, p\}$.

The proof of Theorem 1 is in Appendix A. The variance bound uses Von Neumann's trace inequality, and it can be quite loose when the eigenvalues of C_{xx} decay fast. However, the lower bound and upper bound match when C_{xx} is the identity (the precise non-asymptotic concentration statement is in Lemma 2, Appendix A). Yet in this case, the bias term can go unbounded when $p \to \infty$. The bias lower bound is trivial, but the estimator is trivially unbiased if β is in the span of the data. A more refined lower bound can be computed if one takes into consideration the alignment between the eigenspaces of the data and the eigenspaces of the population covariance (see Theorem 3). The bias and variance upper bounds are mainly based on Koltchinskii and Lounici's concentration inequality [30], similar to Lemma 35 of [3]. Their theorem statements are reproduced in Appendix A for convenience.

The upper bound in Theorem 1 is well-controlled if: 1) The effective rank $r_0(C_{xx}) = \frac{\operatorname{tr}(C_{xx})}{\lambda_1}$ grows slower than n when p increases, so $M \to 0$ as $n \to \infty$, which implies bounded bias and is necessary for bounded variance; 2) The k-th largest eigenvalue λ_k is bounded away from zero and independent of p, and thus λ_1/λ_k is well-conditioned, yielding a small variance. This dimensionless bound shows that PCA-OLS does not exhibit the peaking phenomenon under mild conditions on the population covariance structure.

Tighter risk bounds using spectral gap assumptions

In order to use perturbation analysis to estimate the distance between empirical covariance eigenvectors and their corresponding population covariance eigenvectors, we require a minimum separation among the population covariance eigenvalues. A classical assumption considers a population covariance with possibly repeated eigenvalues, but it requires all empirical covariance eigenvalues corresponding to the same population eigenvalue to be tightly

clustered. Let

$$E := C_{xx} - \frac{1}{n} X^\top X \tag{4.10}$$

be the difference between the empirical and population covariance (i.e., the perturbation). Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the eigenvalues of the population covariance matrix C_{xx}. Let $\Delta_r = \{i : \sigma_i(C_{xx}) = \lambda_r\}$ be the r-th eigenvalue cluster and $m_r := \operatorname{card}(\Delta_r)$ be its multiplicity. Define $g_r := \lambda_r - \lambda_{r+1} > 0, r \geq 1$. Let $\bar{g}_r$ be the spectral gap of eigenvalue λ_r, which is defined as:

$$\bar{g}_r = \begin{cases} g_1 & r = 1 \\ \min(g_{r-1}, g_r) & r \geq 2\,. \end{cases}$$

The assumption used in the analysis [31] asks that the perturbation E is small, in the sense that $\|E\|_{\text{op}} < \frac{\bar{g}_r}{2}$, such that all the empirical eigenvalues $\tilde{\lambda}_j, j \in \Delta_r$ are covered by an interval

$$(\lambda_r - \|E\|_{\text{op}}, \lambda_r + \|E\|_{\text{op}}) \subset (\lambda_r - \bar{g}_r/2, \lambda_r + \bar{g}_r/2)$$

and the rest of the empirical eigenvalues are outside of the interval

$$(\lambda_r - (\bar{g}_r - \|E\|_{\text{op}}), \lambda_r + (\bar{g}_r - \|E\|_{\text{op}})) \supset [\lambda_r - \bar{g}_r/2, \lambda_r + \bar{g}_r/2]\,.$$

To correctly align the leading k clusters of empirical eigenvalues to their population counterparts, the diameter of each population eigenvalue cluster must be strictly smaller than the distance between any two clusters. To this end, we assume

$$\|E\|_{\text{op}} < \frac{1}{4} \min_{1 \leq r \leq k} \bar{g}_r, \tag{4.11}$$

which is the assumption we will use to apply the concentration results from [31] to the top k eigenspace. In addition, we require

$$\operatorname{sgn}(\lambda_i > \lambda_j)\, 2\tilde{\lambda}_i > \operatorname{sgn}(\lambda_i > \lambda_j)\,(\lambda_i + \lambda_j)\,, \forall i \in \{1, \cdots, k\}, j \in \{1, \cdots, p\}, j \neq i, \tag{4.12}$$

which is the condition for the top k eigenspace alignment results in [40].

We remark that the assumptions (4.11) and (4.12) hold when assuming the population spectral gap for the top k eigenspaces, and the sample size n are sufficiently large. From Theorem 9 in [30], there exists a constant c such that for any constant $1 < t < n$, with probability at least $1 - e^{-t}$:

$$\|E\|_{\text{op}} \leq c\lambda_1 \max\left\{\sqrt{\frac{r_0(C_{xx})}{n}}, \frac{r_0(C_{xx})}{n}, \sqrt{\frac{t}{n}}\right\}. \tag{4.13}$$

Therefore if k is fixed, the assumptions may continue to hold even when p is large, as long as the effective rank $r_0(C_{xx}) := \frac{\operatorname{tr}(C_{xx})}{\lambda_1}$ is $o(n)$. Note that these assumption do not hold when C_{xx} is the identity (i.e., isotropic case), nor for the spiked covariance model in [26] (where the top eigenvalue is $\lambda_1 > 1$ and the rest are all 1).

Theorem 2. *Let $x_i \sim \mathcal{N}(0_p, C_{xx})$, $i = 1, \ldots, n$, satisfying spectral gap assumptions (4.11) and (4.12). Let*

$$y_i = x_i^\top \beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. *Let* $w_{ij} = \frac{\lambda_j}{\lambda_i}, k_j^2 = \lambda_j \left(\lambda_j + \operatorname{tr}(C_{xx})\right)$, *and* t *be a constant. Then with probability greater than* $1 - \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{p} \frac{4 w_{ij} k_j^2}{nt(\lambda_i - \lambda_j)^2}$ *we have*

$$\mathbb{V} \leq \frac{\sigma^2}{n} \sum_{i=1}^{k} \left(\frac{\lambda_i}{\lambda_i + \|E\|_{op}} + t \right). \tag{4.14}$$

Furthermore, if we assume k *is fixed and* $n \gg k, n \to \infty$, *with probability greater than any constant* $a \in (0, 1)$ *we have*

$$\mathbb{V} \geq \frac{\sigma^2}{n} \sum_{i=1}^{k} \left(\frac{\lambda_i}{\lambda_i - \|E\|_{op}} - o(1/n) \right). \tag{4.15}$$

The proof of Theorem 2 is in Appendix B. Finally, we produce a lower bound for the bias by treating β as random, which provides an "average-case" analysis.

Theorem 3. *Let* $x_i \sim \mathcal{N}(0_p, C_{xx})$, $i = 1, \ldots, n$, *satisfying spectral gap assumption* (4.12). *Let*

$$y_i = x_i^\top \beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, *and* β *is randomly drawn from an isotropic distribution, where* $\mathbb{E}_\beta[\beta] = 0$, $\mathbb{E}_\beta[\beta\, \beta^\top] = I$. *Let* $k_j^2 = \lambda_j \left(\lambda_j + \operatorname{tr}(C_{xx})\right)$, *and* t *be a constant. Then with probability at least* $1 - \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{p} \frac{4\lambda_j k_j^2}{nt(\lambda_i - \lambda_j)^2}$ *we have*

$$\mathbb{E}_\beta[\mathbb{B}] \geq \sum_{i=k+1}^{p} \lambda_i - kt. \tag{4.16}$$

The proof of Theorem 3 is in Appendix C. Theorem 2 shows that the variance of PCA-OLS depends on the spectral gap for the leading k eigenvalues: larger spectral gap yields better control of the variance. Theorem 3 illustrates that choosing large k decreases the bias on average; it also shows that the larger the spectral gap, the smaller the constant t can be chosen, and thus the tighter the lower bound. As we shall see in Section 5.1, this risk bound has the same leading order term as oracle-PCR.

We remark that for a non-vacuous probability bound, we require $\lambda_i \gg \lambda_j \approx 0$ for $i = 1, \cdots, k$, $j =, k+1, \cdots, p$ (i.e., in the gapped covariance model, or exponential decay model), such that the terms associated with $j = k+1, \cdots, p$ vanish:

$$\sum_{i=1}^{k} \sum_{j\neq i} \frac{4\lambda_j k_j^2}{nt\left(\lambda_i - \lambda_j\right)^2} \approx \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{4\lambda_j^2 \left(\lambda_j + \operatorname{tr}(C_{xx})\right)}{nt\left(\lambda_i - \lambda_j\right)^2}. \tag{4.17}$$

Since k is fixed, the numerator is dominated by $\operatorname{tr}(C_{xx})$, which is $o(n)$ given that $r_0(C_{xx}) = o(n)$ and λ_1 is fixed. Thus, (4.17) tends to 0 (slowly) when $n \to \infty$.

We note that [56] also provides a non-asymptotic upper bound for the risk of PCA-OLS using perturbation analysis. However, they rely on the results from the upper bounds on the excess risk of principal component analysis—the difference between using the empirical eigen-projectors and the population eigen-projectors. In addition to perturbation arguments, we use the notion of effective rank to characterize the data model, which is particularly useful in analyzing the overparameterized setting.

5. Comparison of PCA-OLS with other projection-based methods.

5.1. Oracle-PCR. Xu and Hsu [62] analyzed the double-descent phenomenon of performing principal component regression using the population covariance matrix (instead of the empirical covariance as in PCA-OLS), which we call oracle-PCR. They considered β as random to derive the asymptotic risk. Wu and Xu [59] extended the analysis of oracle-PCR to more general setting of β (i.e., either fixed or random) and its alignment with the eigenvalues of C_{xx}.

We sketch the risk for oracle-PCR, following equation (4.6) by replacing the empirical principal components with their population counterparts. Since the data is generated from a Gaussian distribution, we can assume $C_{xx} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$ without loss of generality. Thus, when C_{xx} is diagonal with $\lambda_1 \geq \cdots \geq \lambda_p$, oracle-PCR essentially truncates the design matrix $X \in \mathbb{R}^{n\times p}$ to its first k columns. In other words, $X\,\Pi_{\text{oracle-PCR}} = X_{[:k]} \in \mathbb{R}^{n\times k}$, where $X_{[:k]}$ denotes the submatrix of the first k columns of X. Let $\hat{\beta}_k = X_{[:k]}^{\dagger}\, Y \in \mathbb{R}^k$ and $\vec{0} \in \mathbb{R}^{p-k}$. The estimator is given by:

$$\hat{\beta}_{\text{oracle-PCR-k}} = [\hat{\beta}_k, \vec{0}] \in \mathbb{R}^p. \tag{5.1}$$

Let $C_{xx,[k:]} \in \mathbb{R}^{(p-k)\times(p-k)}$ be the principal submatrix of C_{xx} by deleting the first k rows and columns. Let $X_{[:k]}$ be the submatrix of the first k columns of X. Recall that $C_{xx} = \sum_{j=1}^p \lambda_j u_j u_j^\top$, and $\beta = C_{xx}^{-1} C_{xy} \equiv \sum_{j=1}^p b_j u_j$. Let the empirical eigenvalues of $\frac{1}{n} X_{[:k]}^\top X_{[:k]}$ be $\tilde{\lambda}'_1 \geq \cdots \geq \tilde{\lambda}'_n$, and $\tilde{u}'_i$ be the empirical eigenvectors $\tilde{\lambda}'_i$. We have

$$\mathcal{R}(\hat{\beta}_{\text{oracle-PCR,k}} \,|\, X) = \beta^\top C_{xx,[k:]}\, \beta + \frac{\sigma^2}{n}\,\mathrm{tr}\left((\frac{1}{n} X_{[:k]}^\top X_{[:k]})^\dagger\, C_{xx}\right) + \sigma^2 \tag{5.2}$$

$$= \sum_{i=k+1}^{p} b_i^2 \lambda_i + \frac{\sigma^2}{n}\,\mathrm{tr}\left(\left(\sum_{i=1}^{k} \frac{1}{\tilde{\lambda}'_i} \tilde{u}'_i \tilde{u}_i'^\top\right)\left(\sum_{j=1}^{p} \lambda_j u_j u_j^\top\right)\right) + \sigma^2. \tag{5.3}$$

Consider the model in Theorem 3 by treating β as random where $\mathbb{E}[b_i^2] = 1$. The bias of oracle-PCR (i.e., first term in (5.3)) has expected value $\sum_{i=k+1}^p \lambda_i$, similar to PCA-OLS (see Theorem 3). The bias also illustrates the alignment between the coefficients of β and the principal components: if they are misaligned in the sense that b_i is large for $i = k+1, \cdots, n$, then the bias is large.

Moreover, oracle-PCR tends to have larger variance than PCA-OLS. Indeed, the submatrix $X_{[:k]} \in \mathbb{R}^{n\times k}$ has smaller empirical singular values than the original data X, $\tilde{\lambda}'_i \leq \tilde{\lambda}_i$ for $i = 1, \cdots, k$. This is a consequence of the interlacing theorem of singular values [47].

However, it is possible that the empirical eigenvectors of oracle-PCR (i.e. $\tilde{u}'_i$) have better alignment with the population eigenvectors (i.e. u_i) than those of PCA-OLS (i.e. $\tilde{u}_i$). Thus, it is not clear whether PCA-OLS always has a smaller variance. Nevertheless, when $k \approx n$, oracle-PCR works with an ill-conditioned $X_{[:k]}$, while PCA-OLS works with a rank-k matrix $X_{\text{PCA,k}}$ that is well-conditioned, given that the effective rank $r_0(C_{xx}) = o(n)$ and the k-th largest population eigenvalue is bounded away from 0 (see Theorem 1).

Remarkably, even when the true population covariance is known, use of the empirical covariance could yield better performance; we demonstrate this in Section 7.2.

5.2. Other projection methods. We briefly discuss a few other projection methods.

Parial Least Squares (PLS): Similar to PCA-OLS, PLS is a data-dependent dimensionality reduction method. Helland and Almøy in [18] compared the asymptotic performance

of PCA-OLS versus PLS in the underparameterized regime where p is fixed and $n \to \infty$. They analyzed the alignment of data features and the regression coefficient via the notion of eigenvalue *relevance*: an eigenvalue is irrelevant for the regression if it corresponds to the principal component that has small correlation with the dependent variable Y. They showed that PCA-OLS performs well when the irrelevant eigenvalues are extremely small or extremely large, while PLS does well for intermediate irrelevant eigenvalue. In the overparamaterized regime, Cook et al. analyzed PLS for various alignments of n, p in the asymptotic regime [12]. They showed that PLS achieves its best performance in data models with many weak features (i.e., abundant regression).

Random orthogonal projection: Another feature extraction method is random orthogonal projection. The recent work [37] provides a very thorough analysis of the risk in this case, interpreting this model as a two-layer linear neural network, where the first layer is a random orthogonal projection (not trained) and the second layer performs ridge regression.

In the isotropic case where C_{xx} is the identity, random orthogonal projections are equivalent to performing oracle-PCR: Since all the population eigenvalues are the same, choosing the first k principal components in oracle-PCR is equivalent to randomly select k orthogonal directions in random orthogonal projection.

Random Gaussian projection:

In [2], Ba et al. analyzed the asymptotic risk of random Gaussian projections in the isotropic covariance case, assuming that both the data X and the projection matrix Π are generated from a Gaussian distribution with zero mean and identity covariance. More quantitative comparisons with PCA-OLS can be found in Supplementary B.

5.3. Data models with increasing number of features and the choice of k. Increasing number of input features can improve generalization for certain data models [3] and methods (such as PLS in abundant regression), but this is not always true. There are some explicit examples, like the one in Section 3 of [23] where an increasing number of features deteriorates the performance.

In practice, the number of features can be made arbitrarily large when fitting the original input features with a flexible functional form, such as a data-independent projection method, or a basis-function expansion (for example, a polynomial or a Fourier series, like the model analyzed in [60] and [20]). In the basis-function expansion setting, each data point y_i is associated with a (scalar or vector) location t_i in some ambient space $\mathbb{R}^p$, and the new features $[X]_{ij}$ are created by basis-function evaluations $g_j(t_i)$. The expansion of the locations t_i into the feature matrix $X \in \mathbb{R}^{n \times k}$ can be viewed as a feature embedding. Hence, the number of functions $g_j(t)$ represents the number of features k; if the basis is infinite, k can be made arbitrarily large. Another popular non-parametric method, Gaussian Processes (GPs), can also have infinite number of parameters; any GP can be seen as the limit as $k \to \infty$ of a specifically created and regularized least-squares, with Mercer's Theorem and the kernel trick converting the k-sums in the outer product $X\,X^\top$ into kernel-function evaluations [34, 20].

Choosing k is an active area of data science research: it appears in the context of projection-based methods, basis-function expansion, and the architectural design of neural networks. In practice, it is typically addressed using cross-validation. The choice of k in PCA-OLS has been studied extensively: from a feature selection perspective, [27] reviewed some iterative procedures to select variables; from a model selection perspective, [21] discussed an improved Bayesian model evidence criteria to select the number of components in a high-dimensional, small sample size setting.

6. Adversarial attacks in linear regression. Adversarial attacks are a very interesting phenomenon that was discovered in the context of deep learning, where imperceptible perturbations of the data can produce huge changes in the predictions of classifiers [54].

Recent work proves that adversarial examples are ubiquitous for classification in the overparameterized regime [5]: Assuming non-zero label noise, the set of adversarial examples is asymptotically dense in the support of data, so there is an adversarial example arbitrarily close to any data point.

In the context of regressions, many kinds of adversarial attacks had been studied as well [8, 32, 33, 24, 46]. In the robust regression literature, most studies rely on the assumption that the feature vector β is sparse [61, 11, 38]. For example, [11] showed that data poisoning attacks can completely change the support of β in sparse regression. Here, we provide straightforward analysis that does not require sparsity assumptions, which is naturally hinted in our discussion of data models with large number of weak features (see Section 5.3).

In what follows, we consider only a "data-poisoning" attack, in which the attacker adds a single data point (x_0, y_0) to the training data with the goal of having it (dramatically) increase the risk on test data.

Definition 1. *(Data-poisoning attack). Let (X, Y) be the original training data. Let us consider an additional adversarial pair (x_0, y_0), and let:*

$$\tilde{X} = \begin{bmatrix} X \\ x_0^\top \end{bmatrix} \in \mathbb{R}^{(n+1)\times p}, \quad \tilde{Y} = \begin{bmatrix} Y \\ y_0 \end{bmatrix} \in \mathbb{R}^{n+1}.$$

The attacker's goal is to maximize the empirical risk subject to the constraints on both x_0 and y_0 (note that x, y might be in different units, so we shall apply the constraint separately).

$$\max_{\|x_0\|\le\epsilon, \|y_0\|\le\epsilon} \|\tilde{Y} - \tilde{X}\beta\|^2. \tag{6.1}$$

We find that, for ordinary (unregularized) least squares, in the overparameterized regime, data-poisoning attacks can be arbitrarily successful. The fundamental reason is that the new p-dimensional data point x_0 can lie very near (but not precisely in) the n-dimensional subspace spanned by the extant data X; it then increases dramatically the condition number of the empirical covariance $X^\top X$ and makes the ordinary least-squares regression arbitrarily sensitive to the training labels Y and y_0. This is in contrast to the underparameterized regime, in which the effect of the attack is limited [33]. The following propositions are very simple. Their proofs are in the Supplementary Material.

Proposition 2. *In the overparameterized regime, ordinary least squares is arbitrarily sensitive to data-poisoning attacks. The risk tends to infinity when the additional adversary feature x_0 is arbitrarily close to the column space of X. In other words, let $\tilde{X} = [X^\top; x_0]^\top$. If $x_0 = \sum_{i=1}^n \alpha_i v_i + \delta$ where $\mathrm{Col}(X) = \mathrm{span}\{v_1, \cdots, v_n\}$, then $\mathcal{R}(\hat{\beta} \,|\, \tilde{X}) \to \infty$ when $\delta \to 0$.*

In contrast, to make successful data-poisoning attack in the underparameterized regime, $X^\top X$ has to be ill-conditioned.

Proposition 3. *In the underparameterized regime, ordinary least squares is arbitrarily sensitive to data-poisoning attack if the smallest eigenvalue of $X^\top X$ is smaller than the attack strength ϵ.*

Corollary 4. *If the k-th largest eigenvalue of $X^\top X$ is much greater than ϵ, then PCA-OLS is robust to data-poisoning attack.*

Proposition 3 and Corollary 4 show that PCA-OLS are robust to data-poisoning attack in the natural PCA setting where the first k eigenvalues have high energy. In other words, data-poisoning attack fails if the empirical covariance of the features is well-conditioned. The same reasoning tells us that ridge regression will be robust to this specific attacks.

We conjecture, and our experiments suggest, that this particular attack will also become bounded for the generative model.

We emphasise that the attacks described in this Section are specifically tailored to exploit the vulnerability of unregularized ordinary least squares. It remains an open problem to study attacks tailored to the other regressions that maximizes the risk. For instance, one could presume an attack for PCA-OLS that imposes a structure in which y is predicted by the low variance components of x. Nevertheless, PCA-OLS can easily detect such attack with an outlier-based defense strategy: If the attack aims to change the k-th largest empirical eigenvalue, then the adversarial pair must lie outside the rank-k principal component subspace, and the magnitude of the attack must grow with n. Thus, the poison point will appear as an outlier compared to the original data features, given that n is large and the population covariance has most energies in the first k components.

Our results also suggest new insights of using PCA as a data preprocessing to defend adversarial attacks in classification settings [7, 10, 1]. [1] showed that the effectiveness of PCA defense depends on choosing the correct number of principal components, in the sense that it matches the intrinsic dimension of the data manifold. This is transparent in our Corollary 4: If PCA-OLS chooses a wrong k that corresponds to a low variance components, the attack becomes considerably powerful.

7. Numerical experiments.

We perform numerical experiments[1] on data generated by the model described in Section 2.

7.1. Generalization and robustness as a function of n, p.

We consider two settings, one in which the number of features p is fixed and the number of samples n vary, and another in which the number of samples is fixed and the number of features vary. In order to define the latter we consider a large $(N+1)\times(N+1)$ covariance matrix Σ_N for $N = 512$ constructed as $WW^\top$ where W is a random matrix with standard i.i.d Gaussian entries. We manipulate the eigenvalues of Σ_N to produce a gap, where the top $k = 32$ eigenvalues are larger than the rest (the manipulation is done by rescaling its top 32 eigenvalues to be 100 times larger). We define C_{xx}^p to be the first $p \times p$ block of Σ_N. Since the eigenvectors of Σ_N are incoherent with the axis, we obtain that C_{xx}^p exhibits the same structure, as illustrated in Figure 7.1 (left).

In Figures 7.2 and 7.1 (right) we report the mean square prediction error (MSE), defined as

$$\mathrm{MSE}(\hat{\beta}, X_{\mathrm{test}}, Y_{\mathrm{test}}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{(x_i^*,y_i^*)\in(X_t^*,Y_t^*)}\frac{1}{n_{\mathrm{test}}}\|x_i^*\hat{\beta}_t - y_i^*\|^2\,, \tag{7.1}$$

where $X_{\mathrm{test}}, Y_{\mathrm{test}}$ denote the test sets with T trials, and each trial (X_t^*, Y_t^*) consists of n_{test} of data samples. We choose $T = 16, n_{\mathrm{test}} = 256$ in our simulations.

For estimators $\hat{\beta}_{\mathrm{OLS}}$, $\hat{\beta}_{\mathrm{generative}}$, $\hat{\beta}_\lambda$ (where the ridge optimal regularization parameter λ is found with cross-validation), and $\hat{\beta}_{\mathrm{PCA},k}$ for $k = 32$. We compare with the prediction error of the true β and with the prediction error of the null estimator $\hat{\beta} = 0$.

In Figure 7.2 we fix number of samples $n = 64$ and vary the number of features p. We observe that OLS exhibits the peaking phenomenon, whereas the other regularized methods do

[1] Code available at: https://github.com/nhuang37/dimensionality_reduction

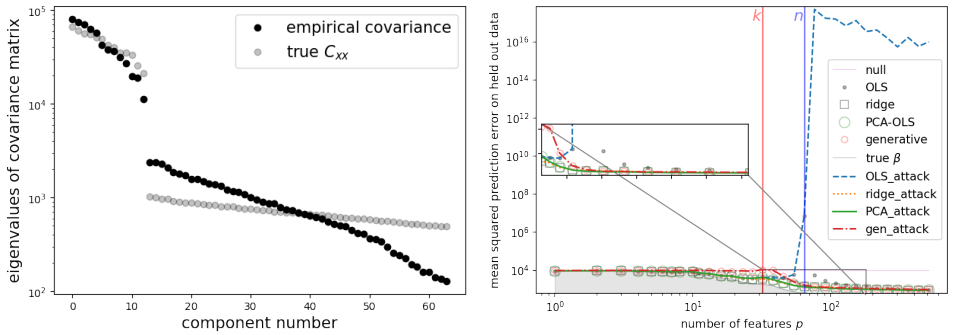

Figure 7.1. ***(Left)*** *Eigenvalues of the empirical and model covariance C_{xx}^p for $p = 64$, with a gap at component number $k = 32$.* ***(Right)*** *MSE for prediction of y on test under data-poisoning attacks of magnitude $\epsilon = 1$, with fixed $n = 64$ and varying p. The attacks in OLS for $p > n$ are arbitrarily successful, while not so effective for other regularized methods.*

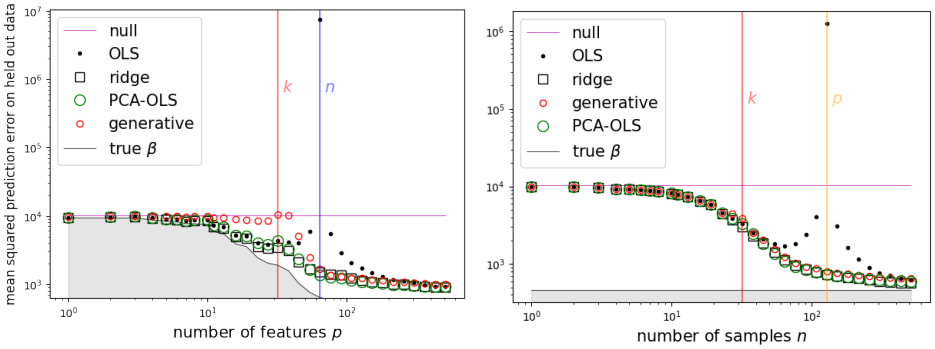

Figure 7.2. ***(Left)*** *MSE for prediction of y on test sets. We consider different regression methods, namely OLS, ridge regression (with optimal ridge parameter found by cross-validation), the generative model described in Section 2, and OLS performed on a PCA projection of the data to dimension $k = 32$. We also report the performance of the null estimator $\hat{\beta} = 0$, and the true linear coefficient β (delimiting the shaded region). In these experiments we fix the number of samples $n = 64$, and we vary the number of features p generating the data according to the model described in Section 7.* ***(Right)*** *MSE for prediction of y with fixed number of features $p = 128$ and varying the number of samples n. We observe that OLS exhibits the peaking phenomenon, whereas the other (regularized) estimators have practically the same, monotonically decreasing risk.*

not. In the overparameterized regime, the risks of all the estimators coincide and consistently decrease with p. The generative model fails to predict the data when the number of features is less than k. This is expected as the latent space has too much freedom, or the latent variable is too powerful such that the model doesn't utilize the signals from the training labels.

In Figure 7.2 we fix the number of features $p = 128$ and vary the number of samples. We observe that ordinary least squares exhibit the peaking phenomenon and the *more data can hurt* behavior at around $p \approx n$, while the rest of the regularized estimators perform similarly, with monotonically decreasing mean square prediction error as the number of samples increases.

In Figure 7.1 (right) we consider the setting from Figure 7.2 (left), and we evaluate the data-poisoning attack described in Section 6 with $\epsilon = 1$. We observe the ordinary least squares estimator is very susceptible to the attack, whereas the regularized methods are robust.

7.2. Comparison of projection-based methods and the choice of k. We compare the performance of five different projection-based methods as a function of the projection dimension k: PCA-OLS; oracle-PCR from [62]; partial least squares from [12]; random Gaussian

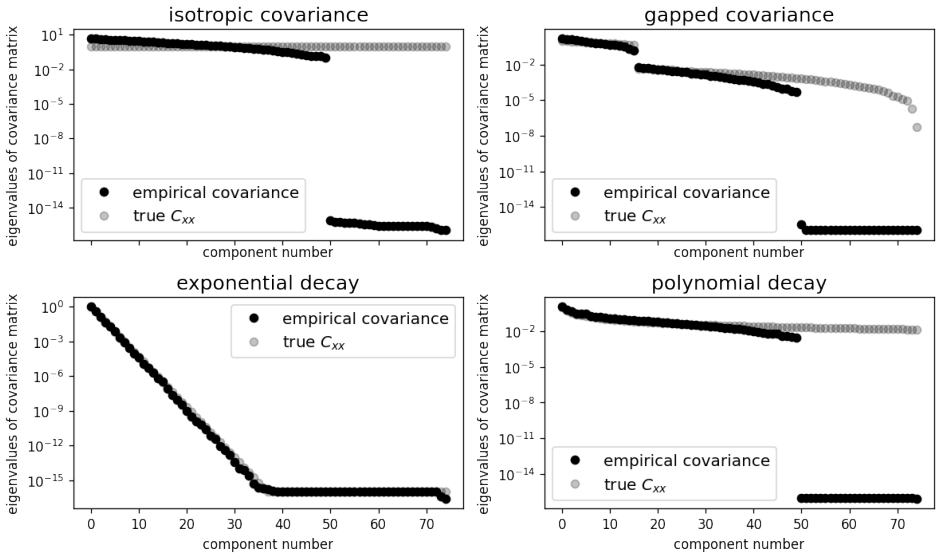

Figure 7.3. *Different covariance structures in our experiments: isotropic covariance refers to C_{xx} being the identity ; gapped covariance refers to a planted eigenvalue gap (at component 16); exponential(polynomial) decay refers to the different decay patterns of the eigenvalues. All the largest eigenvalue is chosen to be* 1.

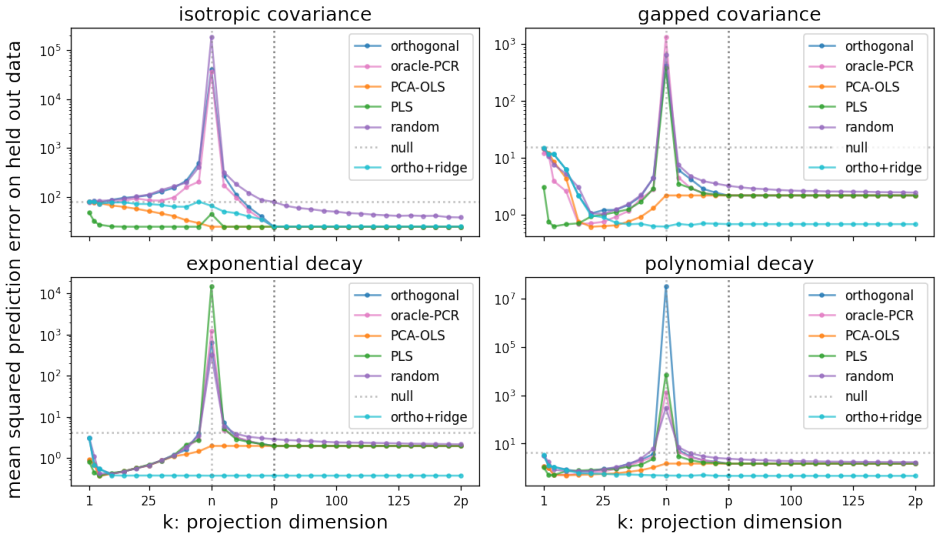

Figure 7.4. *High signal-to-noise (SNR=16) setting with different covariance structure. Random orthogonal projection method is labeled as "orthogonal" for the case without regularization, and "ortho+ridge" for the case with optimally-tuned ridge regularization (by leave-one-out-cross-validation).*

projection from [2]; and orthogonal projection from [37] (including both the unregularized case and optimally-tuned ridge regression case).

To understand the interaction between the method and the data generating process, we consider four different population covariance structures, as illustrated in Figure 7.3. For each covariance model, we fix the number of samples to $n = 50$ and the number of parameters to $p = 75$, and vary the projection dimension k. We generate $\beta \sim \mathcal{N}(0, I_p)$, corresponding to the condition in Theorem 3. The performance is evaluated by out of sample mean square error (equation (7.1)) averaged over $T = 10$ trials with $n_{\text{test}} = 256$. For all random projection methods, we also averaged over five experiments with random weights.

Figure 7.4 illustrates the advantage of *data-dependent* dimensionality reduction, as compared to those overparameterized projection methods that are independent of training data and without regularization, especially when the population covariance exhibits eigenvalue decay. Remarkably, oracle-PCR performs worse than PCA-OLS at almost all choices of k, as analyzed in Section 5.1.

Similarly, random Gaussian projection and random orthogonal projection are inferior to PCA-OLS due to the lack of proper regularization. Nevertheless, random orthogonal projection with ridge regularization achieves similar minimum risk as PCA-OLS, where further overparameterization improves its performance for most cases.

In Supplementary D we provide further numerical experiments where we investigate different signal-to-noise ratio ($\text{SNR} = \frac{\|\beta\|}{\sigma}$) and the setting where the coefficients of β are misaligned with the eigenvalues of C_{xx}.

8. Discussion. Regularization plays an important role in inference. Large-capacity models that can perfectly fit the training data will also generalize well if appropriate regularization is in play [50]. In this article we studied different regression models in the context of the double-descent phenomenon [4], with different forms of regularization. We show that dimensionality reduction is indeed a form of regularization, one that under certain assumptions avoids the "peaking phenomenon" (avoids, in the sense that the risk is bounded when the number of features equals the number of samples). Our difference with previous work in [62] is that our dimensionality reduction is based on the empirical singular values of the features (it is derived from a standard principal components analysis); it is not based on the eigenvalues of the unobservable (true) population covariance.

More precisely, we provide non-asymptotic bounds for the risk of PCA-OLS, which is the ordinary least squares estimator following a projection of the features to their principal components (also known in the literature as PCR or principal component regression). Our main results hold in the overparameterized regime, where the number of samples n is smaller than the number of parameters p, and the effective rank of the data is $o(n)$. A future research direction is to analyze PCA-OLS in the asymptotic regime. A particularly interesting question is under what data conditions can we prove that the risk of PCA-OLS decreases with p in the asymptotic regime $\frac{n}{p} = \gamma < 1$.

Besides analyzing the generalization performance of a particular method, we compare double-descent curves of various projection-based methods. Based on our empirical results, we conjecture that *data-dependent* dimensionality reduction methods are superior to those independent of training data that lack proper regularization. Alternatively, we can view different projection methods as feature selection procedure. For example: PCA-OLS chooses features based on the maximum variance direction in the data; oracle-PCR chooses features with the prior from the true data model; random projection selects feature randomly. Our findings are connected with the discussion in [6] that the double-descent curve depends on the feature selection procedure. Moreover, we show that it is also driven by the data models and regularization.

We also show that unregularized least squares is extremely vulnerable to data-poisoning attacks in the overparameterized regime, whereas other regularized methods are not vulnerable. Our approach has the limitation that the attacks considered were tailored to ordinary least squares and not the regularized methods. We conjecture, however, that it is much harder to achieve arbitrarily large risk when attacking regularized methods.

Many different forms of regularization have similar flavors (for instance l_2, l_1 regularization). We propose a generative model that is closely related to PCA-OLS and PLS as shown

in Section 2. Intuitively, in the overparameterized regime, the generative model learns a low-dimensional latent space that fits the data and labels as closely as possible, while PCA-OLS projects the data into a low-dimensional space that is close to the original data space in terms of the projection error.

Finally, we emphasize that the motivation behind the generative model comes from the physical sciences, where many models used in practice have the generative structure, for instance [44]. In this work, we have proved that its special case, PCA-OLS, does not exhibit the "peaking phenomenon". Our empirical results suggest that this may hold generally for generative models. Analysis of these kinds of models is a good direction for future research.

9. Acknowledgments. We thank Joshua Agterberg, Edgar Dobriban, Liliana Forzani, Daniel Hsu, Carey Priebe, Liza Rebrova, Bernhard Schölkopf and Rachel Ward for relevant discussions. We also thank the anonymous reviewers for giving us constructive feedback that helped us improve this manuscript significantly. SV is partially supported by NSF DMS 2044349, EOARD FA9550-18-1-7007, and the NSF-Simons Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (MoDL) (NSF DMS 2031985).

Appendix A. Proof of Theorem 1.

A key ingredient for the proof of Theorem 1 is the concentration of the eigenvalues of the empirical covariance matrix around the respective eigenvalues of the population covariance matrix. In particular we use the following uniform concentration result by Koltchinskii and Lounici [30].

Theorem 9 in [30]. *Let $x_1, \cdots, x_n$ be the i.i.d. samples from a Gaussian distribution with mean 0 and covariance C_{xx}. Let X be the sample matrix where the i-th row is given by x_i. There exists a constant c such that for any constant $1 < t < n$, with probability at least $1 - e^{-t}$:*

$$\|C_{xx} - \frac{1}{n} X^\top X\|_{op} \leq c\lambda_1 \max\left\{\sqrt{\frac{r_0(C_{xx})}{n}}, \frac{r_0(C_{xx})}{n}, \sqrt{\frac{t}{n}}\right\}, \tag{A.1}$$

where λ_1 is the largest eigenvalue of C_{xx}, and $r_0(C_{xx}) := \frac{\mathrm{tr}(C_{xx})}{\lambda_1}$ is the effective rank.

We first prove Lemmas 1 and 2, and Theorem 1 will follow.

Lemma 1. *Let $E := C_{xx} - \frac{1}{n} X^\top X$. Then*

$$\mathbb{B} \leq \|\beta\|^2 \Big(2\|E\|_{op} + \lambda_{k+1}\Big). \tag{A.2}$$

Proof. Let the spectral decomposition of the empirical covariance matrix be:

$$\frac{1}{n} X^\top X = \sum_{i=1}^{\min\{n,p\}} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top, \qquad X_{PCA,k}^\top X_{PCA,k} = n \sum_{i=1}^{k} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top. \tag{A.3}$$

Thus:

$$\Pi_{X_{\mathrm{PCA}\perp}} = I - (X_{PCA,k}^\top X_{PCA,k})^\dagger X_{PCA,k}^\top X_{PCA,k} = I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top, \tag{A.4}$$

where I is the $p \times p$ identity matrix. The bias term can be bounded by:

$$\mathbb{B} = \mathrm{tr}(\beta^\top \Pi_{X_{\mathrm{PCA}\perp}} C_{xx} \Pi_{X_{\mathrm{PCA}\perp}} \beta) \tag{A.5}$$

$$\leq \|\Pi_{X_{\mathrm{PCA}\perp}} C_{xx} \Pi_{X_{\mathrm{PCA}\perp}}\|_{\mathrm{op}} \|\beta\|^2 \equiv \|\mathbb{C}\|_{\mathrm{op}} \|\beta\|^2, \tag{A.6}$$

where (A.6) follows from Von Neumann's trace inequality. $\|\cdot\|_{\text{op}}$ is the operator-norm for matrices. Note that:

$$\mathbb{C} = (I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top)\, C_{xx}\, (I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top) \tag{A.7}$$

$$= (I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top)\, (C_{xx} - \sum_{i=1}^{k} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top)\, (I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top). \tag{A.8}$$

Equation (A.8) holds because $(I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top)$ and $\sum_{i=1}^{k} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top$ are orthogonal. If $k = p < n$, then $\|I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top\|_{\text{op}} = 0$. So, the bias is trivially 0. Otherwise $\|I - \sum_{i=1}^{k} \tilde{u}_i \tilde{u}_i^\top\|_{\text{op}} = 1$:

$$\|\mathbb{C}\|_{\text{op}} \leq \|C_{xx} - \sum_{i=1}^{k} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top\|_{\text{op}} = \|C_{xx} - \sum_{i=1}^{n} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top + \sum_{i=k+1}^{n} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top\|_{\text{op}} \tag{A.9}$$

$$\leq \|C_{xx} - \frac{1}{n} X^\top X\|_{\text{op}} + \|\sum_{i=k+1}^{n} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^\top\|_{\text{op}} \tag{A.10}$$

$$= \|C_{xx} - \frac{1}{n} X^\top X\|_{\text{op}} + \tilde{\lambda}_{k+1} \tag{A.11}$$

$$\leq 2\|C_{xx} - \frac{1}{n} X^\top X\|_{\text{op}} + \lambda_{k+1}, \tag{A.12}$$

where the last inequality follows from observing that, for any $1 \leq k < n$:

$$\|C_{xx} - \frac{1}{n} X^\top X\|_{\text{op}} = \max_{i=\{1,\cdots,n\}} \left|\lambda_i - \tilde{\lambda}_i\right| \geq \left|\lambda_{k+1} - \tilde{\lambda}_{k+1}\right|. \tag{A.13}$$

■

Lemma 1 shows essentially the same bound that Theorem 4 of [3], with an extra term coming from the $k+1$-th eigenvalue of C_{xx} (that in practice should be small for the data models in which PCA is suitable). Note that (A.10) can be quite loose. It could be refined by applying the Davis-Kahan theorem under additional assumptions on the eigenvalue separation.

Lemma 2. *Let $E := C_{xx} - \frac{1}{n} X^\top X$. Assume $\|E\|_{op} < \lambda_k$, then*

$$\frac{\sigma^2}{n} \frac{k\lambda_p}{\lambda_1 + \|E\|_{op}} \leq \mathbb{V} \leq \frac{\sigma^2}{n} \frac{k\lambda_1}{\lambda_k - \|E\|_{op}}. \tag{A.14}$$

Proof. Recall from (A.3), $\tilde{\lambda}_i$ denotes the i-th eigenvalue of $(1/n)\, X_{\text{PCA}}^\top X_{\text{PCA}}$, for $i = 1, \cdots, k$. λ_i denotes the i-th eigenvalue of C_{xx}. Thus, $1/\tilde{\lambda}_{1+k-i}$ is the i-th largest eigenvalue of $(\frac{1}{n} X_{\text{PCA}}^\top X_{\text{PCA}})^\dagger$. By Von Neumann's trace inequality we have:

$$\text{tr}\left((\frac{1}{n} X_{\text{PCA}}^\top X_{\text{PCA}})^\dagger\, C_{xx}\right) \leq \sum_{i=1}^{k} \frac{\lambda_i}{\tilde{\lambda}_{1+k-i}} \leq \frac{1}{\tilde{\lambda}_k} \sum_{i=1}^{k} \lambda_i. \tag{A.15}$$

By (A.13), $\tilde{\lambda}_k \geq \lambda_k - \|E\|_{\text{op}} > 0$, where the second inequality holds from assumption. So we obtain:

$$\mathbb{V} = \frac{\sigma^2}{n} \text{tr}\left((\frac{1}{n} X_{\text{PCA}}^\top X_{\text{PCA}})^\dagger\, C_{xx}\right) \leq \frac{\sigma^2}{n} \frac{k\lambda_1}{\lambda_k - \|E\|_{\text{op}}}. \tag{A.16}$$

In order to get a lower bound we use Von Neumann's trace inequality:

$$\text{tr}\left((\frac{1}{n}X_{\text{PCA}}^\top X_{\text{PCA}})^\dagger C_{xx}\right) \geq \sum_{i=1}^{k}\frac{\lambda_{p-i+1}}{\tilde{\lambda}_{1+k-i}} \geq \frac{1}{\tilde{\lambda}_1}\sum_{i=1}^{k}\lambda_{p-i+1}. \tag{A.17}$$

By (A.13), $\tilde{\lambda}_1 \leq \lambda_1 + \|E\|_{\text{op}}$, so we obtain:

$$\mathbb{V} = \frac{\sigma^2}{n}\text{tr}\left((\frac{1}{n}X_{\text{PCA}}^\top X_{\text{PCA}})^\dagger C_{xx}\right) \geq \frac{\sigma^2}{n}\frac{k\lambda_p}{\lambda_1 + \|E\|_{\text{op}}}. \tag{A.18}$$

Combining the upper with the lower bound gives us the result. ∎

Proof of Theorem 1. From lemma 1 and lemma 2, we can control the bias and variance simultaneously via $\|E\|_{\text{op}}$. We complete the proof by using the upper bound of $\|E\|_{\text{op}}$ in (A.1).

Appendix B. Proof of Theorem 2 . In order to prove Theorem 2 we use the following results from [40]:

Corollary 4.1 in [40]. *For any weights w_{ij} and real $t > 0$:*

$$\mathbf{P}\left(\sum_{i\neq j} w_{ij}\left\langle \tilde{u}_i, u_j\right\rangle^2 > t\right) \leq \sum_{i\neq j}\frac{4w_{ij}k_j^2}{nt\left(\lambda_i - \lambda_j\right)^2}, \tag{B.1}$$

where $k_j^2 = \lambda_j\left(\lambda_j + \text{tr}(C_{xx})\right)$ for data generated from Gaussian distribution, and $w_{ij} \neq 0$ when $\lambda_i \neq \lambda_j$.

Let $P_r = u_r u_r^\top, \hat{P}_r = \tilde{u}_r\tilde{u}_r^\top$ be the empirical and population eigen-projectors, respectively. We need the following set of concentration results from [31]:

Equation 1.3 in [31].

$$\mathbb{E}\|\hat{P}_r - P_r\|_2^2 = (1 + o(1))\frac{A_r(C_{xx})}{n}, \tag{B.2}$$

where $A_r(C_{xx}) = 2\,\text{tr}\left(P_r C_{xx} P_r\right)\text{tr}\left(C_r C_{xx} C_r\right)$ and the operator C_r is defined as $C_r := \sum_{s\neq r}\frac{P_s}{\mu_r - \mu_s}$.

Equation 1.4 in [31].

$$\text{Var}\left(\|\hat{P}_r - P_r\|_2^2\right) = (1 + o(1))\frac{B_r^2(C_{xx})}{n^2}, \tag{B.3}$$

where $B_r(C_{xx}) := 2\sqrt{2}\,\|P_r C_{xx} P_r\|_2\,\|C_r C_{xx} C_r\|_2$.

Note that under mild assumption, $A_r(C_{xx})$ and $B_r(C_{xx})$ have the same order as the effective rank $r_0(C_{xx})$. Thus, if $r_0(C_{xx}) = o(n)$, then the empirical eigen-projectors concentrate on their population counterparts. This is the crucial assumption to the following asymptotic normality result:

Equation 1.5 in [31]. *Assume effective rank $r_0(C_{xx}) = o(n)$:*

$$\frac{\|\hat{P}_r - P_r\|_2^2 - \mathbb{E}\|\hat{P}_r - P_r\|_2^2}{\text{Var}^{1/2}\left(\|\hat{P}_r - P_r\|_2^2\right)} \sim \mathcal{N}(0,1). \tag{B.4}$$

When stating our concentration results, we often build on the result for the i-th eigenspace, and then use an intersection bound to conclude the probability for the leading k eigenspaces:

Let E_i be the i-th event. By union bound,

$$\mathbf{P}\left(\bigcup_{i=1}^{k} E_i^c\right) \leq \sum_{i=1}^{k} \mathbf{P}\left(E_i^c\right). \tag{B.5}$$

Using De Morgan's Law,

$$\mathbf{P}\left(\left(\bigcap_{i=1}^{k} E_i\right)^c\right) \leq \sum_{i=1}^{k} \left[1 - \mathbf{P}\left(E_i\right)\right] \implies \mathbf{P}\left(\bigcap_{i=1}^{k} E_i\right) \geq \sum_{i=1}^{k} \mathbf{P}(E_i) + 1 - k. \tag{B.6}$$

Proof of Theorem 2. For the upper bound we write

$$\operatorname{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger C_{xx}\right) = \operatorname{tr}\left((\sum_{i=1}^{k} \frac{1}{\tilde{\lambda}_i} \tilde{u}_i \tilde{u}_i^\top)(\sum_{j=1}^{p} \lambda_j u_j u_j^\top)\right) = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{\lambda_j}{\tilde{\lambda}_i} \langle \tilde{u}_i, u_j\rangle^2 \tag{B.7}$$

$$= \sum_{i=1}^{k} \left(\frac{\lambda_i}{\tilde{\lambda}_i}\langle \tilde{u}_i, u_i\rangle^2 + \sum_{j\neq i, j=1}^{p} \frac{\lambda_j}{\tilde{\lambda}_i}\langle \tilde{u}_i, u_j\rangle^2\right) \tag{B.8}$$

$$\leq \sum_{i=1}^{k} \left(\frac{\lambda_i}{\lambda_i + \|E\|_{\mathrm{op}}}\langle \tilde{u}_i, u_i\rangle^2 + \sum_{j\neq i, j=1}^{p} \frac{\lambda_j}{\tilde{\lambda}_i}\langle \tilde{u}_i, u_j\rangle^2\right), \tag{B.9}$$

where the last inequality follows from (A.13). Now, using Corollary 4.1 from [40] (equation (B.1)), for each i, the following event

$$\sum_{j\neq i} \frac{\lambda_j}{\tilde{\lambda}_i}\langle \tilde{u}_i, u_j\rangle^2 \leq t \tag{B.10}$$

holds with probability at least $1 - \sum_{j\neq i} \frac{4w_{ij}k_j^2}{nt(\lambda_i-\lambda_j)^2}$, where $w_{ij} = \frac{\lambda_j}{\tilde{\lambda}_i}, k_j^2 = \lambda_j\left(\lambda_j + \operatorname{tr}(C_{xx})\right)$. Then the probability of all $i = 1, \cdots, k$ terms being upper bounded by t is at least $1 - \sum_{i=1}^{k}\sum_{j\neq i} \frac{4w_{ij}k_j^2}{nt(\lambda_i-\lambda_j)^2}$. Together with the fact that $\langle \tilde{u}_i, u_j\rangle^2 \leq 1$, we have

$$\operatorname{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger C_{xx}\right) \leq \sum_{i=1}^{k} \left(\frac{\lambda_i}{\lambda_i + \|E\|_{\mathrm{op}}} + t\right), \tag{B.11}$$

with probability at least $1 - \sum_{i=1}^{k}\sum_{j\neq i} \frac{4w_{ij}k_j^2}{nt(\lambda_i-\lambda_j)^2}$. Note that this probability is valid when $r_o(C_{xx}) := \frac{\operatorname{tr}(C_{xx})}{\lambda_1} = o(n)$ when k is fixed while $n, p \to \infty$.

For the lower bound we start from equation (B.8) and drop the second term where $j \neq i$:

$$\operatorname{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger C_{xx}\right) \geq \sum_{i=1}^{k} \frac{\lambda_i}{\tilde{\lambda}_i}\langle \tilde{u}_i, u_i\rangle^2 \geq \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i - \|E\|_{\mathrm{op}}}\langle \tilde{u}_i, u_i\rangle^2. \tag{B.12}$$

Let $\Phi(a)$ denote the standard normal distribution. Assume the effective rank $r_0(C_{xx}) = o(n)$ and apply the asymptotic normality result by [30]:

$$\mathbf{P}\left(\frac{\|\hat{P}_r - P_r\|_2^2 - \mathbb{E}\|\hat{P}_r - P_r\|_2^2}{\operatorname{Var}^{1/2}\left(\|\hat{P}_r - P_r\|_2^2\right)} \leq a\right) = \Phi(a). \tag{B.13}$$

Now, observe that

$$\|\hat{P}_r - P_r\|_2^2 = \|\hat{P}_r\|_2^2 + \|P\|_2^2 - 2\langle \hat{P}_r, P_r\rangle = 2 - 2\langle \tilde{u}_i, u_i\rangle^2. \tag{B.14}$$

Thus, with probability $\Phi(a)$:

$$\langle \tilde{u}_i, u_i\rangle^2 \geq 1 - \frac{1}{2}\left(\mathbb{E}\|\hat{P}_r - P_r\|_2^2 + a\,\mathrm{Var}^{1/2}\left(\|\hat{P}_r - P_r\|_2^2\right)\right) = 1 - o(1/n), \tag{B.15}$$

where the expectation and the variance are given in equation (B.2) and (B.3). Recall that both of them have the same order as $r_0(C_{xx})/n$. By assumption, the effective rank $r_0(C_{xx})$ is $o(n)$. Thus, both the expectation and the variance grow as $o(n^{-1})$. Note that throughout our proof of Theorem 2, k is fixed, and thus a is some constant.

Plugging back in equation (B.12) and combining with an intersection bound from equation (B.6), with probability at least $k\Phi(a) + 1 - k$:

$$\mathrm{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger\, C_{xx}\right) \geq \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i - \|E\|_{\mathrm{op}}}(1 - o(1/n)). \tag{B.16}$$

We remark that as $n \to \infty$, there exists a constant a large enough for the probability to be positive. ∎

To summarize, with high probability

$$\frac{\sigma^2}{n}\sum_{i=1}^{k}\left(\frac{\lambda_i}{\lambda_i - \|E\|_{\mathrm{op}}} - o(1/n)\right) \leq \mathbb{V} \leq \frac{\sigma^2}{n}\sum_{i=1}^{k}\left(\frac{\lambda_i}{\lambda_i + \|E\|_{\mathrm{op}}} + t\right). \tag{B.17}$$

Appendix C. Proof of Theorem 3.

Proof. Assume β is randomly drawn from an isotropic distribution: $\mathbb{E}_\beta[\beta] = 0$, $\mathbb{E}_\beta[\beta\,\beta^\top] = I$. Then we provide a lower bound by taking expectation over β:

$$\mathbb{E}_\beta[\mathbb{B}] = \mathbb{E}_\beta\Big[\,\mathrm{tr}\left(\beta^\top\,\Pi_{X_{\mathrm{PCA}\perp}}\,C_{xx}\,\Pi_{X_{\mathrm{PCA}\perp}}\,\beta\right)\Big] = \mathrm{tr}\left(\Pi_{X_{\mathrm{PCA}\perp}}\,C_{xx}\,\Pi_{X_{\mathrm{PCA}\perp}}\,\mathbb{E}_\beta[\beta\,\beta^\top]\right) \tag{C.1}$$

$$= \mathrm{tr}\left(C_{xx}\;(I - \Pi_{X_{\mathrm{PCA}}})\right) = \sum_{j=1}^{p}\lambda_j - \sum_{i=1}^{k}\sum_{j=1}^{p}\lambda_j\langle u_j, \tilde{u}_i\rangle^2 \tag{C.2}$$

$$= \sum_{j=1}^{p}\lambda_j - \sum_{i=1}^{k}\lambda_i\langle u_i, \tilde{u}_i\rangle^2 - \sum_{i=1}^{k}\sum_{j\neq i}^{p}\lambda_j\langle u_j, \tilde{u}_i\rangle^2, \tag{C.3}$$

where the last equation follows by splitting the inner products between population and empirical eigenvectors into terms involving $j = i$ (i.e., large) and $j \neq i$ (i.e., small). The large terms can be bounded by $\langle u_i, \tilde{u}_i\rangle^2 \leq 1$. The small terms can be controlled using Corollary 4.1 from [40] (equation (B.1)), which states the following event E_i

$$\sum_{j\neq i}\lambda_j\langle \tilde{u}_i, u_j\rangle^2 \leq t \tag{C.4}$$

holds with probability at least $1 - \sum_{j\neq i}\frac{4\lambda_j k_j^2}{nt(\lambda_i - \lambda_j)^2}$, where $k_j^2 = \lambda_j\,(\lambda_j + \mathrm{tr}(C_{xx}))$. Then the probability of all $E_i, i = 1, \cdots, k$ being upper bounded by t is at least $1 - \sum_{i=1}^{k}\sum_{j\neq i}\frac{4\lambda_j k_j^2}{nt(\lambda_i - \lambda_j)^2}$. Thus, with probability at least $1 - \sum_{i=1}^{k}\sum_{j\neq i}\frac{4\lambda_j k_j^2}{nt(\lambda_i - \lambda_j)^2}$,

$$\mathbb{E}_\beta[\mathbb{B}] \geq \sum_{j=1}^{p}\lambda_j - \sum_{i=1}^{k}\lambda_i - kt = \sum_{i=k+1}^{p}\lambda_i - kt. \tag{C.5}$$

∎

REFERENCES

[1] S. Alemany and N. Pissinou, *The dilemma between dimensionality reduction and adversarial robustness*, 2020, https://arxiv.org/abs/2006.10885.

[2] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang, *Generalization of two-layer neural networks: An asymptotic viewpoint*, in International Conference on Learning Representations, 2020.

[3] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences, (2020).

[4] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[5] M. Belkin, D. Hsu, and P. Mitra, *Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate*, 2018, https://arxiv.org/abs/1806.05161.

[6] M. Belkin, D. Hsu, and J. Xu, *Two models of double descent for weak features*, SIAM Journal on Mathematics of Data Science, 2 (2020), p. 1167–1180, https://doi.org/10.1137/20m1336072.

[7] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, *Enhancing robustness of machine learning systems via data transformations*, in 2018 52nd Annual Conference on Information Sciences and Systems (CISS), 2018, pp. 1–5, https://doi.org/10.1109/CISS.2018.8362326.

[8] B. Biggio, B. Nelson, and P. Laskov, *Poisoning attacks against support vector machines*, arXiv preprint arXiv:1206.6389, (2012).

[9] A. Canatar, B. Bordelon, and C. Pehlevan, *Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks*, (2020), https://arxiv.org/abs/2006.13198.

[10] N. Carlini and D. Wagner, *Adversarial examples are not easily detected: Bypassing ten detection methods*, in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3–14.

[11] Y. Chen, C. Caramanis, and S. Mannor, *Robust sparse regression under adversarial corruption*, in International Conference on Machine Learning, PMLR, 2013, pp. 774–782.

[12] R. D. Cook, L. Forzani, et al., *Partial least squares prediction in high-dimensional regression*, The Annals of Statistics, 47 (2019), pp. 884–908.

[13] S. d'Ascoli, L. Sagun, and G. Biroli, *Triple descent and the two kinds of overfitting: Where & why do they appear?*, arXiv preprint arXiv:2006.03509, (2020).

[14] P. S. Dhillon, D. P. Foster, S. M. Kakade, and L. H. Ungar, *A risk comparison of ordinary least squares vs ridge regression*, The Journal of Machine Learning Research, 14 (2013), pp. 1505–1511.

[15] L. E. Frank and J. H. Friedman, *A statistical view of some chemometrics regression tools*, Technometrics, 35 (1993), pp. 109–135.

[16] M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, *Jamming transition as a paradigm to understand the loss landscape of deep neural networks*, Physical Review E, 100 (2019), p. 012115.

[17] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, 2020, https://arxiv.org/abs/1903.08560.

[18] I. S. Helland and T. Almøy, *Comparison of prediction methods when only a few components are relevant*, Journal of the American Statistical Association, 89 (1994), pp. 583–591.

[19] M. Henmi and S. Eguchi, *A paradox concerning nuisance parameters and projected estimating functions*, Biometrika, 91 (2004), pp. 929–941.

[20] D. W. Hogg and S. Villar, *Fitting very flexible models: Linear regression with large numbers of parameters*, arXiv preprint arXiv:2101.07256, (2021).

[21] D. C. Hoyle, *Automatic pca dimension selection for high dimensional data and small sample sizes.*, Journal of Machine Learning Research, 9 (2008).

[22] A. K. Jain and B. Chandrasekaran, *39 dimensionality and sample size considerations in pattern recognition practice*, Handbook of statistics, 2 (1982), pp. 835–855.

[23] A. K. Jain, R. P. W. Duin, and J. Mao, *Statistical pattern recognition: A review*, IEEE Transactions on pattern analysis and machine intelligence, 22 (2000), pp. 4–37.

[24] A. Javanmard, M. Soltanolkotabi, and H. Hassani, *Precise tradeoffs in adversarial training for linear regression*, 2020, https://arxiv.org/abs/2002.10477.

[25] T. Jebara, *Machine Learning: Discriminative and Generative*, The Springer International Series in Engineering and Computer Science, Springer US, 2012.

[26] I. M. Johnstone and D. Paul, *Pca in high dimensions: An orientation*, Proceedings of the IEEE, 106 (2018), pp. 1277–1292.

[27] I. T. Jolliffe, *Principal components in regression analysis*, in Principal component analysis, Springer, 1986, pp. 129–155.
[28] S. d. Jong, B. M. Wise, and N. L. Ricker, *Canonical partial least squares and continuum power regression*, Journal of Chemometrics: A Journal of the Chemometrics Society, 15 (2001), pp. 85–100.
[29] P. Ju, X. Lin, and J. Liu, *Overfitting can be harmless for basis pursuit: Only to a degree*, 2020, https://arxiv.org/abs/2002.00492.
[30] V. Koltchinskii, K. Lounici, et al., *Concentration inequalities and moment bounds for sample covariance operators*, Bernoulli, 23 (2017), pp. 110–133.
[31] V. Koltchinskii, K. Lounici, et al., *Normal approximation and concentration of spectral projectors of sample covariance*, Annals of Statistics, 45 (2017), pp. 121–157.
[32] F. Li, L. Lai, and S. Cui, *On the adversarial robustness of subspace learning*, IEEE Transactions on Signal Processing, 68 (2020), p. 1470–1483, https://doi.org/10.1109/tsp.2020.2974676.
[33] F. Li, L. Lai, and S. Cui, *Optimal feature manipulation attacks against linear regression*, arXiv preprint arXiv:2003.00177, (2020).
[34] W. Li, *Generalization error of minimum weighted norm and kernel interpolation*, SIAM Journal on Mathematics of Data Science, 3 (2021), pp. 414–438.
[35] Y. Li, J. Bradshaw, and Y. Sharma, *Are generative classifiers more robust to adversarial attacks?*, in International Conference on Machine Learning, PMLR, 2019, pp. 3804–3814.
[36] Z. Li, C. Xie, and Q. Wang, *Provable more data hurt in high dimensional least squares estimator*, 2020, https://arxiv.org/abs/2008.06296.
[37] L. Lin and E. Dobriban, *What causes the test error? going beyond bias-variance via anova*, (2020), https://arxiv.org/abs/2010.05170.
[38] C. Liu, B. Li, Y. Vorobeychik, and A. Oprea, *Robust linear regression against training data poisoning*, in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, New York, NY, USA, 2017, Association for Computing Machinery, p. 91–102, https://doi.org/10.1145/3128572.3140447.
[39] F. Liu, Z. Liao, and J. A. K. Suykens, *Kernel regression in high dimension: Refined analysis beyond double descent*, (2020), https://arxiv.org/abs/2010.02681.
[40] A. Loukas, *How close are the eigenvectors of the sample and actual covariance matrices?*, in International Conference on Machine Learning, 2017, pp. 2228–2237.
[41] W. F. Massy, *Principal components regression in exploratory statistical research*, Journal of the American Statistical Association, 60 (1965), pp. 234–256.
[42] S. Mei and X. Zhu, *Using machine teaching to identify optimal training-set attacks on machine learners*, in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, AAAI Press, 2015, p. 2871–2877.
[43] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, *Optimal regularization can mitigate double descent*, arXiv preprint arXiv:2003.01897, (2020).
[44] M. Ness, D. W. Hogg, H.-W. Rix, A. Y. Ho, and G. Zasowski, *The cannon: A data-driven approach to stellar label determination*, The Astrophysical Journal, 808 (2015), p. 16.
[45] A. Y. Ng and M. I. Jordan, *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes*, in Advances in neural information processing systems, 2002, pp. 841–848.
[46] D. L. Pimentel-Alarcón, A. Biswas, and C. R. Solís-Lemus, *Adversarial principal component analysis*, in 2017 IEEE International Symposium on Information Theory (ISIT), 2017, pp. 2363–2367.
[47] J. F. Queiró, *On the interlacing property for singular values and eigenvalues*, Linear Algebra and Its Applications, 97 (1987), pp. 23–28.
[48] C. E. Rasmussen, *Gaussian processes in machine learning*, in Summer School on Machine Learning, Springer, 2003, pp. 63–71.
[49] R. Rosipal and N. Krämer, *Overview and recent advances in partial least squares*, in International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", Springer, 2005, pp. 34–51.
[50] J. Rougier and C. E. Priebe, *The exact form of the 'ockham factor'in model selection*, The American Statistician, (2020), pp. 1–16.
[51] M. Slawski et al., *On principal components regression, random projections, and column subsampling*, Electronic Journal of Statistics, 12 (2018), pp. 3673–3712.
[52] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, *A jamming transition from under-to over-parametrization affects loss landscape and generalization*, arXiv preprint arXiv:1810.09665, (2018).
[53] M. Stone and R. J. Brooks, *Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression*, Journal

of the Royal Statistical Society: Series B (Methodological), 52 (1990), pp. 237–258.

[54] C. Szegedy, W. Zaremba, I. Sutskever, J. B. Estrach, D. Erhan, I. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, in 2nd International Conference on Learning Representations, ICLR 2014, 2014.

[55] T. Tarpey, R. T. Ogden, E. Petkova, and R. Christensen, *A paradoxical result in estimating regression coefficients*, The American Statistician, 68 (2014), pp. 271–276.

[56] M. Wahl, *A note on the prediction error of principal component regression*, 2019, https://arxiv.org/abs/1811.02998.

[57] A. G. Wilson and P. Izmailov, *Bayesian deep learning and a probabilistic perspective of generalization*, 2020, https://arxiv.org/abs/2002.08791.

[58] H. Wold, *Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares*, in Evaluation of econometric models, Elsevier, 1980, pp. 47–74.

[59] D. Wu and J. Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, 2020, https://arxiv.org/abs/2006.05800.

[60] Y. Xie, R. Ward, H. Rauhut, and H.-H. Chou, *Weighted optimization: better generalization by smoother interpolation*, arXiv preprint arXiv:2006.08495, (2020).

[61] H. Xu, C. Caramanis, and S. Sanghavi, *Robust pca via outlier pursuit*, IEEE transactions on information theory, 58 (2012), pp. 3047–3064.

[62] J. Xu and D. J. Hsu, *On the number of variables to use in principal component regression*, in Advances in Neural Information Processing Systems, 2019, pp. 5094–5103.

[63] F. Yang, S. Liu, E. Dobriban, and D. P. Woodruff, *How to reduce dimension with pca and random projections?*, arXiv preprint arXiv:2005.00511, (2020).

Supplementary material.

Supplementary A. Bias-variance decomposition of the risk. In order to analyze the risk of the PCA-OLS estimator $\hat{\beta} = \hat{\beta}_{\mathrm{PCA},k}$ we use the standard decomposition of bias plus variance. In this particular case it is

$$\tilde{\beta} = \Pi_{X_{\mathrm{PCA}}}\,\beta \tag{A.1}$$

$$\mathcal{R}(\hat{\beta}\,|\,X) = \underbrace{\mathbb{E}_{x_*}[(x_*^\top(\beta-\tilde{\beta}))^2\,|\,X]}_{\text{bias squared}} + \underbrace{\mathbb{E}_{Y,x_*}[(x_*^\top(\hat{\beta}-\tilde{\beta}))^2\,|\,X]}_{\text{variance}} + \sigma^2 \tag{A.2}$$

$$= \underbrace{\beta^\top \Pi_{X_{\mathrm{PCA}\perp}} C_{xx}\,\Pi_{X_{\mathrm{PCA}\perp}}\,\beta}_{\text{bias squared}} + \underbrace{\frac{\sigma^2}{n}\operatorname{tr}\left((\frac{1}{n}X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger\, C_{xx}\right)}_{\text{variance}} + \sigma^2, \tag{A.3}$$

where X_{PCA}, $\Pi_{X_{\mathrm{PCA}}}$, and $\Pi_{X_{\mathrm{PCA}\perp}}$ are the equivalents of their non-PCA counterparts for the rank-k PCA approximation to X.

Proof. Let $X_p := X_{\mathrm{PCA},k}$. We have:

$$\hat{\beta} = X_p^\dagger\, Y, \tag{A.4}$$

$$Y = X\beta + \epsilon = (X_p + X_p^\perp)\beta + \epsilon, \tag{A.5}$$

$$\mathcal{R}(\hat{\beta}\,|\,X) = \mathbb{E}_{Y,x_*}[(x_*^\top(\beta-\hat{\beta}))^2\,|\,X] + \sigma^2 \tag{A.6}$$

$$= \mathbb{E}_{Y,x_*}[(x_*^\top(\beta - X_p^\dagger\,(X\beta+\epsilon)))^2\,|\,X] + \sigma^2 \tag{A.7}$$

$$= \mathbb{E}_{Y,x_*}[(x_*^\top(I - X_p^\dagger X)\,\beta - x_*^\top X_p^\dagger\,\epsilon)^2\,|\,X] + \sigma^2 \tag{A.8}$$

$$= \mathbb{E}_{Y,x_*}[(x_*^\top(I - X_p^\dagger X_p)\,\beta)^2\,|\,X] + \mathbb{E}_{Y,x_*}[(x_*^\top X_p^\dagger\,\epsilon))^2\,|\,X] + \sigma^2 \tag{A.9}$$

$$= \beta^\top \Pi_{X_{p\perp}} C_{xx}\,\Pi_{X_{p\perp}}\beta + \operatorname{tr}(X_p^{\dagger\top}\,\Sigma\,X_p^\dagger\,\mathbb{E}\left[\epsilon\epsilon^\top|X\right]) + \sigma^2 \tag{A.10}$$

$$= \beta^\top \Pi_{X_{p\perp}} C_{xx}\,\Pi_{X_{p\perp}}\beta + \sigma^2\operatorname{tr}(X_p^\dagger\,X_p^{\dagger\top}\,\Sigma) + \sigma^2 \tag{A.11}$$

$$= \beta^\top \Pi_{X_{p\perp}} C_{xx}\,\Pi_{X_{p\perp}}\beta + \frac{\sigma^2}{n}\operatorname{tr}\left((\frac{1}{n}X_p^\top X_p)^\dagger\, C_{xx}\right) + \sigma^2. \tag{A.12}$$

Note that the cross term in (A.9) does vanish since ϵ has zero mean conditioned on X, and is independent of x^*. Note that (A.11) follows from the assumption that noise is i.i.d. Finally, (A.12) follows from the fact that:

$$X_p^\dagger\,X_p^{\dagger\top} = (X_p^\top X_p)^\dagger, \tag{A.13}$$

which can be obtained by letting $X_p = \sum_{i=1}^k s_i\tilde{v}_i\tilde{u}_i^\top$ and observing $X_p^\dagger = \sum_{i=1}^k \frac{1}{s_i}\tilde{u}_i\tilde{v}_i^\top$. ∎

Supplementary B. Random Gaussian projections on isotropic data. Given the data matrix X with rows $x_i \sim \mathcal{N}(0_p, I_p)$, $i = 1,\ldots,n$ and a random Gaussian projection matrix $\Pi = [w_1,\cdots,w_k] \in \mathbb{R}^{p\times k}$, where $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0_p, p^{-1}I_p\right)$, [2] investigated three cases and concluded the following:

- Case 1: $\operatorname{rank}(X\,\Pi) = p < \min\{n,k\}$: this effectively reduced to the underparameterized case studied in [17]. In the limit of $n,p,k\to\infty$, the bias tends to zero and the variance tends to $\frac{p}{n-p}\sigma^2$.
- Case 2: $\operatorname{rank}(X\,\Pi) = k < n < p$: this is similar to PCA-OLS in the overparameterized regime. The bias is no longer zero, due to the dimensionality reduction, while the variance tends to $\frac{k}{n-k}\sigma^2$.

- Case 3: $\operatorname{rank}(X\,\Pi) = n < k < p$: in this overparameterized setting, the random projection lifts up the features to a higher-dimensional space. The risk decreases monotonically with k.

We focus on comparing PCA-OLS with random Gaussian projection in the overparameterized setting where $p > n$ (case 2, 3).

Let $\hat{\beta}_R$ be the random Gaussian projection method estimator, which is computed by $\hat{\beta}_R = \Pi(X\Pi)^\dagger Y$. Let $\|\beta\|^2 = r^2, \operatorname{Var}(\epsilon) = \sigma^2, \gamma_1 = p/n, \gamma_2 = k/n$.

Case 2: When $k < n$, By Theorem 1 in [2],

$$\mathcal{R}(\hat{\beta}_R \mid X) \to \frac{\gamma_1 - \gamma_2}{\gamma_1|\gamma_2 - 1|} r^2 + \frac{\gamma_2}{|\gamma_2 - 1|}\sigma^2 = \frac{p-k}{p(1-k/n)} r^2 + \frac{k}{n-k}\sigma^2. \tag{B.1}$$

From the bias upper bound of PCA-OLS in isotropic setting ([56] section 3.2.1) and our variance bound in Theorem 1, with high probability, there exists a constant $C > 1$ such that:

$$\mathcal{R}(\hat{\beta}_{\text{PCA}} \mid X) \le \frac{p-k}{p} r^2 + C\frac{k}{n}\sigma^2. \tag{B.2}$$

Comparing (B.1) and (B.2), PCA-OLS has a smaller bias and potentially a smaller variance than random Gaussian projections.

Case 3: If $\gamma_2 \to \infty$, by Theorem 1 in [2],

$$\mathcal{R}(\hat{\beta}_R \mid X) \to \frac{\gamma_2|\gamma_1 - 1|}{\gamma_1|\gamma_2 - 1|} r^2 + \frac{|\gamma_1 - 1| + |\gamma_2 - 1|}{|\gamma_1 - 1|\,|\gamma_2 - 1|}\sigma^2 \tag{B.3}$$

$$\overset{\gamma_2 \to \infty}{\to} \frac{|\gamma_1 - 1|}{\gamma_1} r^2 + \frac{1}{|\gamma_1 - 1|}\sigma^2 \tag{B.4}$$

$$= \mathcal{R}(\hat{\beta}_{\text{OLS}} \mid X). \tag{B.5}$$

where the last equality follows from Theorem 2 in [17]. This shows that the risk of $\hat{\beta}_R$ is equivalent to the risk of OLS estimator. However, $\mathcal{R}(\hat{\beta}_R \mid X)$ monotonically decreases with k (on both bias and variance term). This shows $\hat{\beta}_R$ is strictly worse than OLS (i.e, PCA-OLS) for all k when $k > n$, regardless of the signal-to-noise ratios.

We remark that there are many variants of random Gaussian projection. For example, [51] showed that under stronger assumptions (i.e. the random projections are Johnson-Lindenstrauss transforms), the performance of random Gaussian projections are of the same order as PCA-OLS.

To conclude, the random Gaussian projections in [2] serve as a theoretical tool to analyze the double-descent risk curve, while in practice, stronger assumption are needed to improve its performance as a preprocessing method for regression.

Supplementary C. Adversarial attacks.

Proof of Proposition 2. Under the overparameterized regime where $p > n$, the $\hat{\beta}_{OLS}$ is given by equation (2.6). Thus:

$$\hat{\beta}_{poison} = \tilde{X}^\top(\tilde{X}\tilde{X}^\top)^{-1}\tilde{Y} \tag{C.1}$$

$$= \tilde{X}^\top \begin{bmatrix} XX^\top & Xx_0 \\ x_0^\top X^\top & x_0^\top x_0 \end{bmatrix}^{-1} \tilde{Y} \tag{C.2}$$

$$= \tilde{X}^\top \begin{bmatrix} f_1(\frac{1}{h}) & f_2(\frac{1}{h}) \\ f_3(\frac{1}{h}) & \frac{1}{h} \end{bmatrix} \tilde{Y}, \tag{C.3}$$

$$h := x_0^\top (I - X^\top(XX^\top)^{-1}X)x_0, \tag{C.4}$$

where h is the square of the projection of x_0 onto the p-dimensional space orthogonal to the span of X (i.e., the null space of $X^\top$), f_1, f_2, f_3 are linear functions in $\frac{1}{h}$. We assume the block matrix in equation (C.2) is invertible and thus $h \neq 0$.

Now, we choose

$$x_0 = \frac{\epsilon}{\|\sum_{i=1}^n \alpha_i v_i + \delta\|}\Big(\sum_{i=1}^n \alpha_i v_i + \delta\Big), \tag{C.5}$$

such that $\|x_0\| \leq \epsilon$. Let $\delta \to 0$, then (C.4) becomes

$$h = x_0^\top (I x_0 - X^\top (XX^\top)^{-1} X x_0) \to x_0^\top (x_0 - x_0) = 0. \tag{C.6}$$

■

Thus, h is arbitrarily small and the risk grows to infinity, and the attack is immensely successful.

In practice, we choose x_0 be a (random) linear combination of the columns of X plus a small noise, and then normalize it to have $\|x_0\| = \epsilon$. Meanwhile, y_0 can be chosen arbitrarily as x_0 drives the success of the attack.

Proof of Proposition 3. In the underparameterized regime ($p < n$), the data-poisoning attack becomes:

$$\hat{\beta}_{\text{poison}} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} \tag{C.7}$$

$$= (X^\top X + x_0 x_0^\top)^{-1} \tilde{X}^\top \tilde{Y}. \tag{C.8}$$

Here, $X^\top X$ is full rank, in contrast to the low rank matrix $XX^\top$ in the overparameterized setting. The attack effectively adds a rank-1 matrix $x_0 x_0^\top$ with $\|x_0\|^2 \leq \epsilon^2$. If the smallest eigenvalue of $X^\top X$ is less than ϵ, then the attack can push the smallest eigenvalue of $X^\top X + x_0 x_0^\top$ infinitely close to 0, making the risk of OLS grow arbitrarily. ■

Proof (sketch) of Corolary 4. PCA-OLS first projects the features $X \in \mathbb{R}^{n \times p}$ to a low-dimensional space $\mathbb{R}^{n \times k}$, and then perform OLS on a rank-k approximation of the features, $X_{\text{PCA},k}$. Given $k < \min\{n, p\}$, PCA-OLS is effectively in the underparameterized regime. The smallest eigenvalue of $X_{\text{PCA},k}$ is the k-th largest eigenvalue of $X^\top X$. Thus, if $\lambda_k(X^\top X) \gg \epsilon$, the attack has minimal effect in changing the smallest eigenvalue of $X_{\text{PCA},k}$, and the risk of PCA-OLS under attack will not deviate much from the original risk. ■

Supplementary D. Further numerical experiments.

In a low SNR setting (Figure S1), all the methods perform worse, as expected. In particular, the performance of PLS deteriorates drastically with larger k, under the isotropic covariance model and gapped covariance model. This suggests the higher sensitivity of PLS to the signal-to-noise effect. In comparison, PCA-OLS seems to be more robust, as the shape of its risk curves does not change significantly.

In Figure S2 we study the impact of the alignment of the signal of β and the principal components in C_{xx}, by letting $\beta = [1, 2, \cdots, p-1, p]$. In other words, large coefficients of β are aligned with principal components of C_{xx} with small eigenvalues. In this case, we correct the SNR ratio to keep the same noise variance σ^2 as in the high SNR setting (shown in Figure 7.4). In this misalignment setting, the results for the gapped covariance model changed the most, compared to true β that distributes even weights to all PCs (i.e., Figure 7.4). As shown in Figure S2, PCA-OLS achieves the lowest MSE at $k = n$. A much larger k is needed as the signals concentrate in the principal components with small eigenvalues. Comparatively,

PLS reaches the lowest MSE at a smaller k than PCA-OLS. This illuminates the choice of k depends on both the data covariance as well as its alignment with signals on β.

Similar to [59], we also observe that the risk of oracle-PCR decreases with k in the over-parameterized regime, for both the aligned and misaligned settings.

Finally, we analyze the bias and variance terms for different methods in Figure S3 [2]. The bias term is computed by:

$$\mathbb{B} = \beta^\top \Pi_{X_{p\perp}} C_{xx} \Pi_{X_{p\perp}} \beta, \tag{D.1}$$

where $\Pi_{X_{p\perp}} = I - \Pi(X\Pi)^\dagger X$, following the derivation in Supplementary A. Then we compute the variance by subtracting $\mathbb{B}$ and σ^2 from MSE. Note that this is only an approximation of the true bias and variance component (as the MSE is averaged over Monte-carlo simulations, not the exact risk). As shown in Figure S3, for most cases: the bias-variance trade-off appears for $k < n$; while both bias and variance monotonically decrease for $k > n$. Note that the bias of PCA-OLS is large with small k for the isotropic covariance model (but even larger for other projection methods except PLS). On the other hand, with eigenvalue decays (row 2-4), PCA-OLS achieves low bias, and does not suffer from high variance.

[2] Excluding the ridge regularized orthogonal projection method, as the standard bias variance decomposition of OLS does not apply.

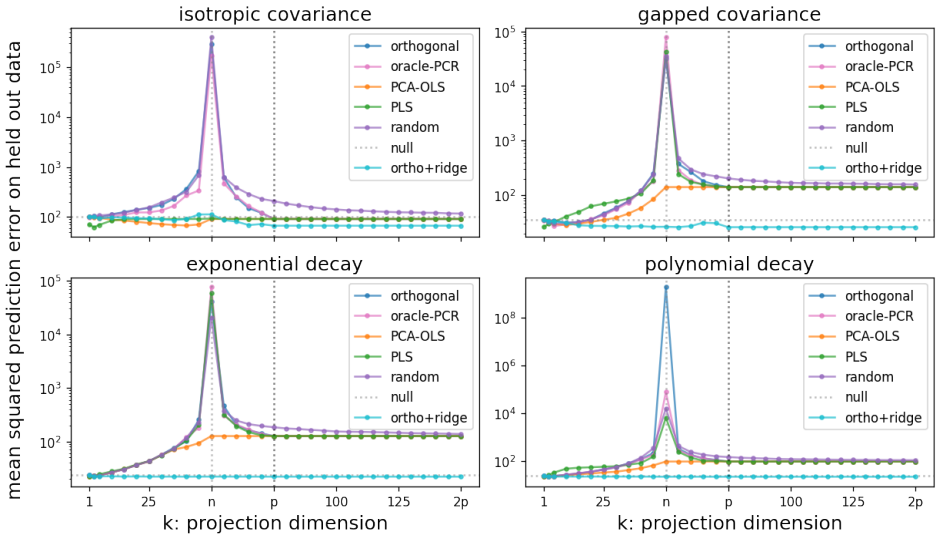

Figure S1. *Low signal-to-noise (SNR=2) setting with different covariance structures.*

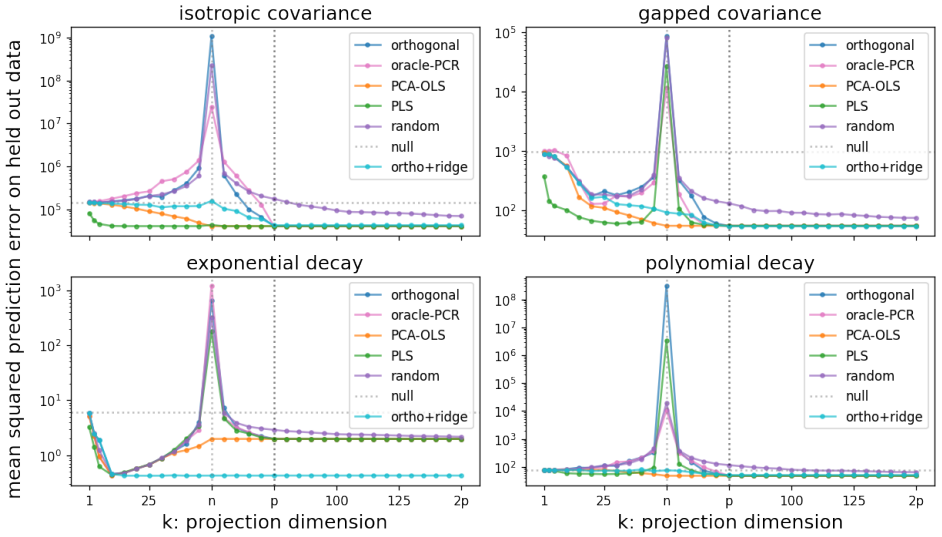

Figure S2. *Misalignment setting with the coefficient of β is inversely related to the eigenvalues in C_{xx}.*

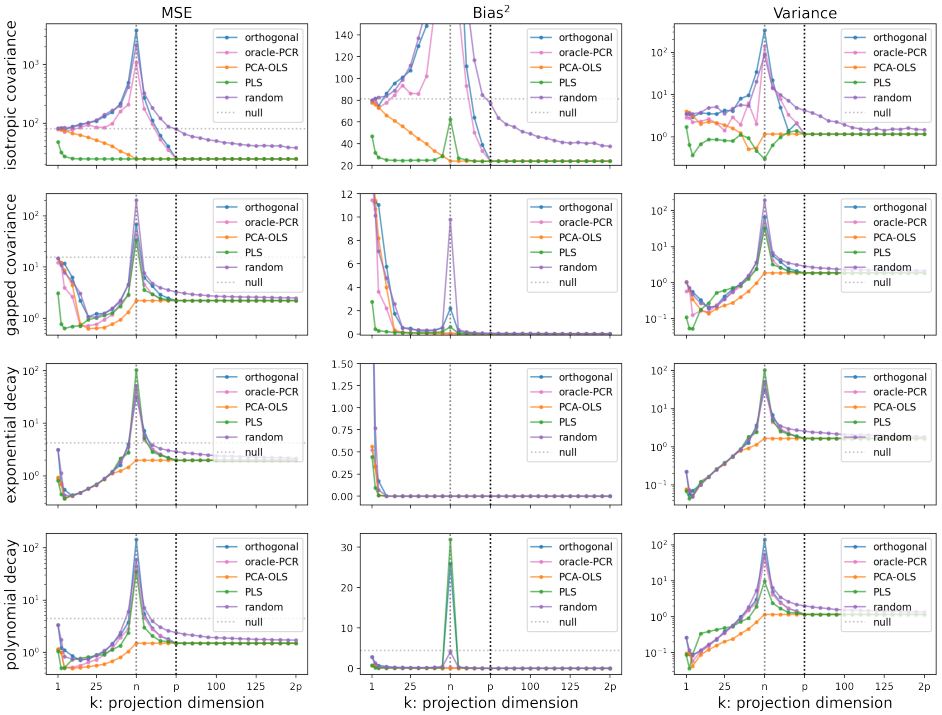

Figure S3. *High signal-to-noise (SNR=16) setting with different covariance structures and bias-variance decomposition for selected methods.*