No Double Descent in Principal Component Regression: A High-Dimensional Analysis

Daniel Gedon [1] **Antônio H. Ribeiro** [1] **Thomas B. Schön** [1]

Abstract

Understanding the generalization properties of large-scale models necessitates incorporating realistic data assumptions into the analysis. Therefore, we consider Principal Component Regression (PCR)—combining principal component analysis and linear regression—on data from a low-dimensional manifold. We present an analysis of PCR when the data is sampled from a spiked covariance model, obtaining fundamental asymptotic guarantees for the generalization risk of this model. Our analysis is based on random matrix theory and allows us to provide guarantees for high-dimensional data. We additionally present an analysis of the distribution shift between training and test data. The results allow us to disentangle the effects of (1) the number of parameters, (2) the data-generating model and, (3) model misspecification on the generalization risk. The use of PCR effectively regularizes the model and prevents the interpolation peak of the double descent. Our theoretical findings are empirically validated in simulation, demonstrating their practical relevance.

1. Introduction

The study of overparameterized models with more features than training data points offers a natural route to gain theoretical understanding when it comes to the successes of large-scale models with good generalization properties (Neyshabur et al., 2015; Zhang et al., 2017). The observation that the generalization error often decreases in the overparameterized regime and the framing as 'double descent' (Belkin et al., 2019) boosted research in this direction even if generalization of large models was already studied before (Bartlett & Mendelson, 2002; Dziugaite & Roy, 2017; Belkin et al., 2018; Advani et al., 2020).

Principal Component Regression (PCR) is a widely adopted model where the input data is projected onto the principal components of the data and then a linear model is fitted to the projected data (Pearson, 1901; Jolliffe, 2002). It is a simple but effective model that is used in many real-world applications for its interpretable regularization effect. Examples include exploratory statistical research (Massy, 1965), econometrics (Geweke, 1996), genetics (Wang & Abbott, 2008), robotics (Vijayakumar & Schaal, 2000) and many more. These real-world datasets are often high-dimensional but lie on a low-dimensional manifold (Tenenbaum et al., 2000). In this work, we therefore consider the case where PCR is applied to data that is sampled from a spiked covariance model (Johnstone, 2001) which is a widely adopted model for high-dimensional data (Baik et al., 2005; Baik & Silverstein, 2006). The spiked covariance model assumes that the data is sampled from a low-dimensional subspace. Thus, the feature covariance matrix is given by a base covariance $\boldsymbol{C}_0$ representing noise and some spikes for the d-dimensional data subspace, yielding $\boldsymbol{C} = \boldsymbol{C}_0 + \sum_{i=1}^{d} \boldsymbol{v}_i \lambda_i \boldsymbol{v}_i^\top$.

While the double descent is well studied for linear regression models (Bartlett et al., 2020; Hastie et al., 2022; Mei & Montanari, 2022), the effect of the double descent for PCR is not well understood. To visualize PCR under high-dimensional inputs for a real-world data example, we use the Diverse MAGIC wheat data set (Scott et al., 2021). Here, we subsample the genotypes uniformly for a varying number of features p while keeping the number of samples n fixed. Figure 1 shows the risk of PCR and full linear regression on this data set. While in this example there is no reason to assume that the data is sampled from a spiked covariance model which is a linear data generator, we can observe that (1) linear regression has a double descent curve and (2) PCR effectively regularizes the model and for sufficiently many principal components, the risk approaches the linear regression risk for small and large number of features p. Therefore, even though our analysis focuses on the linear case, it can describe the qualitative behaviour of PCR on real-world data.

[1] Department of Information Technology, Uppsala University, Sweden. Correspondence to: Daniel Gedon <daniel.gedon@it.uu.se>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

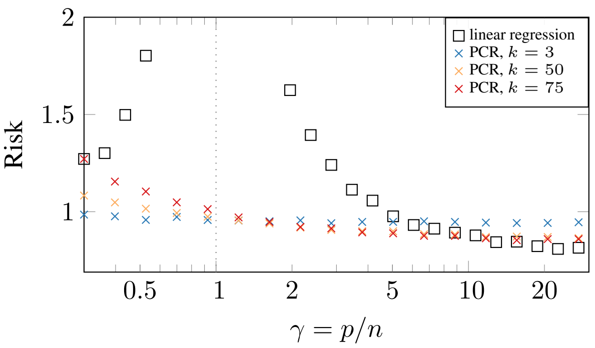


Figure 1: **Risk on real-world data.** PCR and full regression on diverse MAGIC wheat genetics data set.

Our analysis is based on random matrix theory that allows us to provide asymptotic guarantees for high-dimensional data. Random matrix theory is a well-established tool in the analysis of overparameterized models (Pennington & Worah, 2017; Hastie et al., 2022). We apply its results from the spiked covariance model to the generalization risk of PCR. Therefore, we extend the results of unsupervised feature analysis from PCA to the supervised learning setting of PCR.

Our main contribution is an analysis of the *asymptotic generalization risk of PCR* on data sampled from a spiked covariance model. We show the connection of the risk of PCR to the number of parameters, data assumptions and model misspecification. We further provide an analysis of the *distribution shift* between training and test data, a scenario that is often encountered in practice. Our theoretical findings are empirically *validated through simulation*.

Naturally, we expect that PCA with an appropriately chosen number of principal components k regularises the model effectively and prohibits high risk. Our analysis thus provides insights into the precise mechanisms that control the risk and yields fundamental guarantees to rely on for practitioners.

2. Background and problem

Throughout the paper we use bold capital letters $\boldsymbol{X}$ to denote matrices, bold lower-case letters $\boldsymbol{x}$ for vectors, and lower-case letters x for scalars. The identity matrix of size p is denoted by $\boldsymbol{I}_p$. Estimated values are denoted by hats, e.g. $\hat{\boldsymbol{C}}$ is the estimate of the covariance matrix $\boldsymbol{C}$.

2.1. Data generating process.

Let the eigendecomposition of a covariance matrix be $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ is a matrix of sorted eigenvalues and $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_p]$ is a matrix of eigenvectors. The singular value decomposition (SVD) of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with n samples and p features is denoted by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{S} = \operatorname{diag}(s_1, \dots, s_p)$ is a matrix of sorted singular values and $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_p]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_p]$ are matrices of left and right singular vectors, respectively.

Take a base covariance $\boldsymbol{C}_0$ and a low-rank perturbation covariance $\boldsymbol{C}_z$ with d spiked eigenvalues $\lambda_1, \dots, \lambda_d$ and corresponding eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$. The *spiked covariance model* is then defined by $\boldsymbol{C} = \boldsymbol{C}_0 + \operatorname{diag}(\boldsymbol{C}_z, \boldsymbol{0})$ (Johnstone, 2001).

We assume that the base covariance is $\boldsymbol{C}_0 = \boldsymbol{I}_p$ and the eigenvectors $\boldsymbol{v}_i$ are the canonical basis vectors $\boldsymbol{e}_i$. Furthermore, we let the eigenvalues be $\lambda_i = \exp(-i\alpha)$, which describes an exponentially decaying spectrum with decay rate $\alpha \geq 0$. Our analysis does not require this eigenvalue decay or a specific rate α. However, we consider this spectrum in our experiments because fast-decaying eigenvalues occur in many real-world examples. As the data-generating model, we consider the latent factor model which connects the spiked covariance model to regression outcomes.

Definition 1. The *latent factor model* is the linear model $\boldsymbol{x}_i = \boldsymbol{W} r_w \boldsymbol{z}_i + \boldsymbol{e}_i$, and $y_i = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \varepsilon_i$. With latent variables $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_z)$, with diagonal covariance $\boldsymbol{C}_z$, feature noise $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, outcome noise $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, feature matrix $\boldsymbol{W} \in \mathbb{R}^{p \times d}$ such that $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_d$. Let $r_w^2 = \frac{p}{\operatorname{Tr}(\boldsymbol{\Lambda}_z)} \rho_x$ to control the feature signal-to-noise-ratio (SNR) $\rho_x = \frac{\mathbb{E}\left[\|\boldsymbol{W} r_w \boldsymbol{z}\|_2^2\right]}{\mathbb{E}\left[\|\boldsymbol{e}\|_2^2\right]}$, label noise $\varepsilon_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_\varepsilon^2)$ and let $\boldsymbol{\theta} = 1_d \frac{r_\theta}{\sqrt{\operatorname{Tr}(\boldsymbol{\Lambda}_z)}}$ to control the outcome SNR $\rho_y = \frac{\mathbb{E}\left[\|\boldsymbol{\theta}^\top \boldsymbol{z}\|_2^2\right]}{\mathbb{E}\left[\|\varepsilon\|_2^2\right]} = \frac{r_\theta^2}{\sigma_\varepsilon^2}$.

This ensures the population covariance follows the spiked covariance as $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^\top\right] = \boldsymbol{C}$. With this definition $\boldsymbol{C} = \boldsymbol{\Lambda}$ as the population eigenvectors are the canonical basis vectors $\boldsymbol{V} = \boldsymbol{I}_p$. Thus, the feature matrix equals the first d population eigenvectors $\boldsymbol{W} = \boldsymbol{V}_d$.

This low-rank latent factor model can equivalently be expressed as a linear model directly between features and outcome by $y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \nu_i$. With $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$, $\boldsymbol{\beta} = r_w \boldsymbol{W} \boldsymbol{C}_z (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{\theta}$, and $\sigma_\nu^2 = \sigma_\varepsilon^2 + \boldsymbol{\theta}^\top (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{C}_z \boldsymbol{\theta}$. For details, see Appendix C.1. The latent factor model and equivalent direct model are depicted in Figure 2 (left) as the data generating models.

2.2. Random matrix theory background

Let $\hat{\lambda}_i$ denote the eigenvalues of the sample covariance matrix $\hat{\boldsymbol{C}} = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}$, and denote the empirical measure of eigenvalues $\hat{\mu}_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{\hat{\lambda}_i}$ where δ is the Dirac measure. We use results from random matrix theory (Tao, 2023;

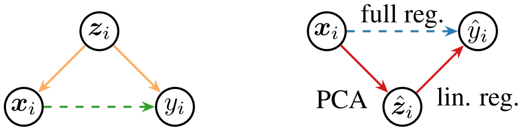


Figure 2: **Model definitions.** *Left:* Data-generating models: Latent factor model (orange) and equivalent direct model (green). *Right:* Estimation models: PCR (red) and full regression model (blue).

Anderson et al., 2010) which allow us to state asymptotic distributions to which the empirical measure converges.

Marčenko-Pastur. With $d = 0$ spikes we get the isotropic case. The empirical measure converges to the Marčenko-Pastur distribution (Marčenko & Pastur, 1967). Informally, the Marčenko-Pastur distribution states that the sample eigenvalues concentrate close to the value 1 for small $\gamma := \frac{p}{n}$ and spread for large γ.

> **Theorem 1** (Marčenko & Pastur (1967)). Let $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0, \infty)$ and sample $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p)$. Then, almost surely the empircal measure $\hat{\mu}_p$ converges weakly to the Marčenko-Pastur distribution with density $f(x)$
>
> $$\begin{cases} f^{MP}(x) = \frac{1}{2\pi\gamma x}\sqrt{(\gamma_+ - x)(x - \gamma_-)}, & \gamma \leq 1 \\ F(dx) = (1 - 1/\gamma)\delta_0(dx) + f^{MP}(x)dx, & \gamma > 1, \end{cases}$$
>
> with δ_0 as the unit point mass at 0, upper and lower boundaries of $f^{MP}(x)$ as $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$.

Eigenvalue shift. Let $d = 1$ spikes, i.e. $\lambda_1 > 1$. The distribution of the sample eigenvalue $\hat{\lambda}_1$ changes with λ_1, transitioning at the critical point $1 + \sqrt{\gamma}$. While generally the bulk of the distribution stays Marčenko-Pastur, there are two cases of interest for the spike:

- $\lambda_1 \in [1, 1 + \sqrt{\gamma}]$: The spike follows limiting Tracy-Widom $n^{2/3}\frac{\hat{\lambda}_1 - \mu(\gamma)}{\sigma(\gamma)} \xrightarrow{\mathcal{D}} TW_1$, with $\mu(\gamma) = (1 + \sqrt{\gamma})^2$ and $\sigma(\gamma) = (1 + \sqrt{\gamma})^{4/3}\gamma^{-1/6}$ (Johnstone, 2001; Baik et al., 2005; Bloemendal & Virág, 2013).
- $\lambda_1 > 1 + \sqrt{\gamma}$: The spike follows a Normal $n^{1/2}\frac{\hat{\lambda}_1 - \mu(\lambda,\gamma)}{\sigma(\lambda,\gamma)} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$, with $\mu(\lambda, \gamma) = \gamma\frac{\lambda}{\lambda - 1} + \lambda$ and $\sigma^2(\lambda, \gamma) = 2\lambda^2(1 - \frac{\gamma}{(\lambda-1)^2})$ (Baik & Silverstein, 2006; Paul, 2007; Yang & Johnstone, 2018).

This highlights an upward shift of the spike sample distribution mean. Hence, the sample eigenvalue $\hat{\lambda}_1$ will separate from the bulk of the Marčenko-Pastur distribution for $\lambda_1 > 1 + \sqrt{\gamma}$. See Figure 3 for an illustrative example.

This is extendable to $d > 1$ with a spike multiplicity of one (Baik et al., 2005; Paul, 2007). For $d > 1$ spikes with multiplicity one, we assume that the sample eigenvalues will be distributed according to the spiked covariance model distribution with a bulk of Marčenko-Pastur and d normally distributed spikes. Let the fraction of population eigenvalues above the critical point be $\frac{d}{p} \to \phi \in (0, 1)$, then the mass of the distribution related to the spike is given by ϕ.

Stieltjes transform. Following Anderson et al. (2010) we define the Stieljes transform of a measure.

> **Definition 2** (Stieltjes (1894)). Let μ be a positive, finite measure on $\mathbb{R}$, then the *Stieltjes transform* of the measure is given by $\varphi_\mu(z) := \int_{\mathbb{R}} \frac{\mu(d\sigma)}{\sigma - z}$ with $z \in \mathbb{C}\backslash\mathbb{R}_+$.

The utility of the Stieltjes transform is that if for a sequence of measures $\{\mu_1, \mu_2, \dots\}$ the Stieltjes transform converges pointwise $\varphi_{\mu_p} \to \varphi_\mu$, then $\mu_p \to \mu$ weakly, see Anderson et al. (2010). Moreover, the Stieltjes transform of the empirical measure $\widehat{\mu}_p$ is

$$\varphi_{\widehat{\mu}_p}(z) = \frac{1}{p}\operatorname{Tr}[(\hat{\boldsymbol{C}} - z\boldsymbol{I}_p)^{-1}]. \tag{1}$$

The Stieltjes transform can be used to prove different limiting distributions for the empirical measure (Bai & Silverstein, 2010) or Bach (2023) for an application-oriented presentation of these results. Here, we provide numerical verification for the results above of the eigenvalue shift but with $d > 1$: In Figure 4 we illustrate that $\varphi_{\widehat{\mu}_p}(z)$ and the Stieltjes transform of the distribution it converges to $\varphi_\mu(z)$ are close for large values of p.

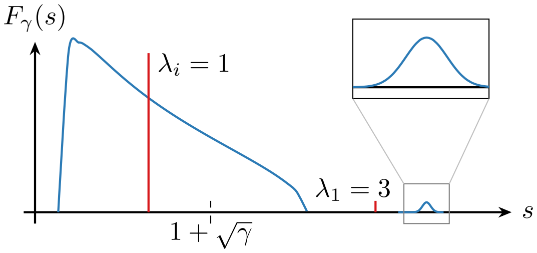


Figure 3: **Eigenvalue spectrum for spiked covariance model.** Sample distribution of $d = 1$ spike (blue) of population eigenvalues (red, $\lambda_1 = 3$, others $\lambda_i = 1$) for $\gamma = 0.3$, $n = 500$. The spike has a normally distributed sample eigenvalue with $\mu = 3.45$.

Eigenvector shift. In the spiked covariance model, the top-d eigenvectors can be inconsistent. Let $\boldsymbol{v}_i \in \mathbb{R}^p$ be unit population eigenvectors with sample eigenvectors $\hat{\boldsymbol{v}}_i \in \mathbb{R}^p$.

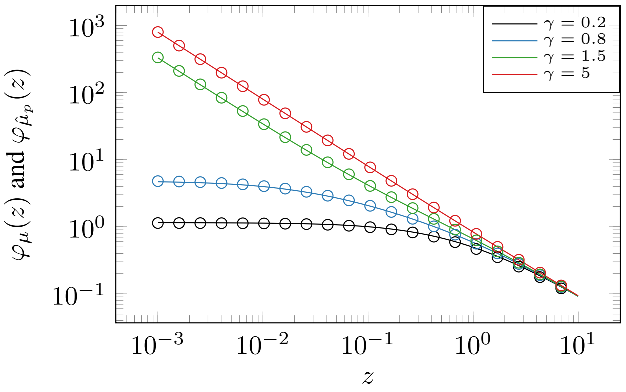


Figure 4: **Stielties transform $\varphi(z)$ of eigenvalue distribution.** Solid lines denote numerical evaluation of the Stieltjes transform Definition 2; '○' markers denote the empirical transform (1). We use $d = 10$ spikes here.

As $p/n \to \gamma$,

$$(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2 \to \begin{cases} \frac{1-\gamma/(\lambda_i-1)^2}{1+\gamma/(\lambda_i+1)} & \text{for } \lambda_i > 1+\sqrt{\gamma} \\ 0 & \text{for } \lambda_i \in [1, 1+\sqrt{\gamma}]. \end{cases} \qquad (2)$$

This implies that the angle between population $\boldsymbol{v}_i$ and sample eigenvectors $\hat{\boldsymbol{v}}_i$ grows as the spike eigenvalue λ_i decreases. Eigenvectors below the critical point resemble isotropic, randomly placed vectors within a hypersphere (Paul, 2007; Johnstone et al., 2009). An overview of high-dimensional PCA is given in Johnstone & Paul (2018).

2.3. Problem formulation

As the data generator, we use the latent factor model and as the estimation model, we use PCR, both of which are visualized in Figure 2. As a baseline estimation model, we choose the unregularized full regression model. We consider the case where the number of spikes d is fixed but unknown. This implies that the signal is concentrated on a low-dimensional (linear) manifold. With fixed d, we analyse the effect of the number of features by presenting results for the risk in the low-dimensional $\gamma < 1$ and high-dimensional $\gamma > 1$ regime.

PCR model. Let $\hat{\boldsymbol{C}}$ be the sample covariance matrix and $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_k]$ be the matrix of sample eigenvectors truncated to the first $k \leq p$ principal components. The *PCR model* is defined as the linear model $\hat{y}_i = \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{z}}_i$ with $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}})^+ \hat{\boldsymbol{Z}}^\top \boldsymbol{y}$ and $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{V}}$.

Full regression model. Let the linear model $\hat{y}_i = \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i$ with $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{y}$ be the unregularized least squares estimator. We denote this as the *full regression model*[1] to contrast it with the PCR model. Note that the full regression model is not low-rank. Furthermore, it is a special case of PCR with $k = p$ and orthogonal features.

Risk. Let $\hat{\boldsymbol{\theta}}$ be a parameter estimator. Then, the *risk* is the mean squared error of the predictions $R(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{(\boldsymbol{x}_0, y_0)\sim\mathcal{D}}\left[(y_0 - \hat{y}_0(\boldsymbol{x}_0))^2\right]$, with data distribution $\mathcal{D}$.

3. Analysis of in-distribution risk

3.1. Risk of PCR

The PCR model jointly estimates SVD components and parameters $\boldsymbol{\theta}$. We first choose the number of principal components k for truncation in the SVD and then estimate the parameters $\hat{\boldsymbol{\theta}}$. Hence, let $\hat{\boldsymbol{S}}$, $\hat{\boldsymbol{U}}$, $\hat{\boldsymbol{V}}$ be the first k singular values and left, right singular vectors, respectively, then the least squares estimator of $\hat{\boldsymbol{\theta}}$ on the latent space projection $\boldsymbol{Z}$ is given by

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + \hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \boldsymbol{\nu}. \qquad (3)$$

To compute the expected risk of PCR, we let $\boldsymbol{\Pi} = \boldsymbol{I}_p - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top$ be the projection matrix onto the subspace orthogonal to the principal components of $\hat{\boldsymbol{C}}$. Then, we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\theta}})\right] =& \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta} \\ &+ \frac{\sigma_\nu^2}{n}\operatorname{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}}\right) + \sigma_\nu^2. \end{aligned} \qquad (4)$$

The terms can be interpreted as squared bias, variance and irreducible error. The result highlights that the variance term contains a projection of the covariance onto its estimated k-dimensional subspace.

For asymptotic results, we generalize the eigenvector shift in the spiked covariance model from the expression $(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2$ in (2) to products of eigenvectors $(\boldsymbol{V}^\top \hat{\boldsymbol{V}})^2$ where we have cross products with $i \neq j$. Let $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_k]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]$ be the estimated and true eigenvectors truncated to the first $k \leq p$ principal components, respectively. Define $\hat{\boldsymbol{V}}_d \in \mathbb{R}^{d\times k}$ and $\boldsymbol{V}_{d,d} \in \mathbb{R}^{d\times d}$ as the matrices where the eigenvectors are truncated to the first d eigenvector elements. Then, the eigenvector shift product is

$$\boldsymbol{P}_k = \begin{cases} \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \ldots, (\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k)^2, 0, \ldots, 0\right), & k < d, \\ \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \ldots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2\right), & k = d, \\ \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2 + c_1^2, \ldots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2 + c_d^2\right), & k > d, \end{cases} \qquad (5)$$

with the correction factor $c_i^2 = (k-d)\frac{1-(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2}{p-d}$ for $k > d$. Note that $\boldsymbol{P}_k$ is a diagonal matrix with entries depending

[1] We use $\boldsymbol{\beta}$ when modelling the interaction between $\boldsymbol{x}$ and y directly and we use $\boldsymbol{\theta}$ when modelling the interaction of a latent variable with the outcome, e.g. $\boldsymbol{z}$ and y.

on the choice of k. The main part in (5) is that off-diagonal values are asymptotically equivalent to zero and simplifies the following analysis. The proofs for (3)-(5) and all following theorems are in Appendix C.

Asymptotic PCR risk. With the generalized eigenvector shift, we can find asymptotic expressions for the risk (4). The *asymptotic squared bias* term is given by

$$\begin{aligned}\mathrm{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 = \bar{\boldsymbol{\beta}}^\top \big(&\boldsymbol{\Lambda}_d - \boldsymbol{\Lambda}_d \boldsymbol{P}_k - \boldsymbol{P}_k \boldsymbol{\Lambda}_d + \boldsymbol{P}_k \\ &+ \boldsymbol{P}_k r_w^2 \boldsymbol{C}_z \boldsymbol{P}_k\big) \bar{\boldsymbol{\beta}}\end{aligned} \tag{6}$$

with $\bar{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta} = r_w \boldsymbol{C}_z (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{\theta}$ and $\boldsymbol{\Lambda}_d \in \mathbb{R}^{d \times d}$ as truncation of the population eigenvalue matrix $\boldsymbol{\Lambda}$ to the first d dimensions. The *asymptotic variance* term is

$$\begin{aligned}\mathrm{Var}_\gamma(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n} \Bigg(&\mathrm{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k) \frac{1}{\mu(\boldsymbol{\Lambda}, \gamma)} \right] \\ &+ (p-d) \int_{s_c}^{(1+\sqrt{\gamma})^2} \frac{1}{s} dF_\gamma(s) \Bigg)\end{aligned} \tag{7}$$

with $\mu(\boldsymbol{\Lambda}, \gamma)$ as diagonal matrix with entries $\mu(\lambda_i, \gamma)$ as mean of the spike eigenvalue distribution, F_γ as the Marčenko-Pastur distribution and s_c the value in $\mathbb{R}$ which satisfies $\max\left(\frac{k-d}{p-d}, 0\right) = \int_{s_c}^{(1+\sqrt{\gamma})^2} dF_\gamma(s)$. Combining both terms yields the following theorem for the asymptotic risk of PCR.

Theorem 2 (Asymptotic PCR risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0, \infty)$, the expected risk of PCR will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}} \left[R(\hat{\boldsymbol{\theta}}) \right] \to \mathrm{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 + \mathrm{Var}_\gamma(\hat{\boldsymbol{\theta}}) + \sigma_\nu^2.$$

The theorem implies that the squared bias is a scaled version of the d-dimensional subspace of the eigenvalues $\boldsymbol{\Lambda}$. Moreover, in the variance, we see that for $k \leq d$ we have that $s_c = (1 + \sqrt{\gamma})^2$ meaning that the integral term is zero. Hence, the variance is a scaled version of the inverse mean of the spike eigenvalue distribution $\mu(\lambda_i, \gamma)$. The results hold for eigenvalues below and above the phase transition threshold.

In linear regression, the interpolation peak at $\gamma = 1$ originates from an increasing variance value (Hastie et al., 2022). For PCR, Theorem 2 and (7) highlight that the second term for the variance integrates the spikes of the sample distribution, see Figure 3. This term is only included if $k > p$ that is governed by s_c. Thus, choosing k determines which parts of the data distribution, i.e. the spiking components and the Marčenko-Pastur, are considered. Appropriately choosing k therefore helps to consider only the spiking components and disregard components from the Marčenko-Pasturdistribution that represent noise.

3.2. Risk of baseline methods

As a reference, we state the null predictor, i.e. $\hat{\boldsymbol{\theta}} = 0$. In this case, the expected null risk becomes $\mathbb{E}_{\boldsymbol{\nu}} \left[R(\hat{\boldsymbol{\theta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{C} \boldsymbol{\beta} + \sigma_\nu^2$, which has zero variance and contains an unprojected squared bias term. Let us restate the full regression risk from (Hastie et al., 2022, Lemma 1), which is a special case of (4) for $k = p$

$$\mathbb{E}_{\boldsymbol{\nu}} \left[R(\hat{\boldsymbol{\beta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi} \boldsymbol{C} \boldsymbol{\Pi} \boldsymbol{\beta} + \frac{\sigma_\nu^2}{n} \mathrm{Tr}(\boldsymbol{C} \hat{\boldsymbol{C}}^{-1}) + \sigma_\nu^2. \tag{8}$$

Below we have the asymptotic result for full regression with the spiked covariance model. Here, the bias is zero for $\gamma < 1$ while the variance term remains unchanged from the asymptotic PCR result. This result is a special case of the PCR result in Theorem 2 for $k = p$.

Theorem 3 (Asymptotic full regression risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0, \infty)$, the expected risk of the full regression model will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}} \left[R(\hat{\boldsymbol{\beta}}) \right] \to \mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2 + \mathrm{Var}_\gamma(\hat{\boldsymbol{\beta}}) + \sigma_\nu^2$$

with the asymptotic squared bias term as $\mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2 = 0$ for $\gamma < 1$ and $\mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2$ as in Theorem 2 with $k = p$ for $\gamma \geq 1$; and the variance term $\mathrm{Var}_\gamma(\hat{\boldsymbol{\beta}})$ equal to the definition of the variance in Theorem 2 with $k = p$.

4. Analysis of covariate-shifted risk

In this section, we change the data generator from Definition 1 to one inspired by Emami et al. (2020) to include covariate shift. Specifically, we introduce a shift in the latent factors between training and test time. Let us assume that the eigenvectors $\boldsymbol{V}$ of the training (or source) covariance $\boldsymbol{C}_S = \boldsymbol{V} \boldsymbol{\Lambda}_S \boldsymbol{V}^\top$ and test covariance $\boldsymbol{C}_T = \boldsymbol{V} \boldsymbol{\Lambda}_T \boldsymbol{V}^\top$ are the same but the eigenvalues $\boldsymbol{\Lambda}_S$, $\boldsymbol{\Lambda}_T$ differ. This covariate shift relates to scenarios where the underlying structure remains consistent, represented by the unchanged eigenvectors. The variations in eigenvalues reflect real-world situations where the magnitude of certain factors changes without altering their relationship.

Definition 3. Let the covariance matrices of the latent factors be $\boldsymbol{C}_{z,S} = \mathrm{diag}\,(\lambda_{1,S}, \ldots, \lambda_{d,S})$ and $\boldsymbol{C}_{z,T} = \mathrm{diag}\,(\lambda_{1,T}, \ldots, \lambda_{d,T})$ for the training and test data, respectively. Then, we have $\boldsymbol{z}_S \sim \mathcal{N}(0, \boldsymbol{C}_{z,S})$, and $\boldsymbol{x}_{i,S} = \boldsymbol{W} r_w \boldsymbol{z}_{i,S} + \boldsymbol{e}_{i,S}$, and $y_{i,S} = \boldsymbol{\theta}^\top \boldsymbol{z}_{i,S} + \varepsilon_{i,S}$ for the training data and similarly for the test data.

Note that we assume the number of spikes d between training and test data distribution is equal. Furthermore, while it is not necessary for the theory, in the experiments in Section 5.2 we connect training and test data by defining a correlation between their eigenvectors of the covariances $\boldsymbol{C}_{z,S}$ and $\boldsymbol{C}_{z,T}$.

4.1. Risk of PCR

We use the parameter estimator (3) but write it as $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + \hat{\boldsymbol{S}}_S^{-1}\hat{\boldsymbol{U}}^\top \boldsymbol{\nu}$, with $\hat{\boldsymbol{S}}_S$ as the training data singular values to highlight the explicit dependence on the training data. Also, we update the definition for the test risk under covariate shift as $R(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{(\boldsymbol{x}_0,y_0)\sim\mathcal{D}_T}\left[(y_0 - \hat{y}_0(\boldsymbol{x}_0))^2\right]$ with $\mathcal{D}_T$ as the test data distribution.

We can generalize the PCR risk from (4) under covariate shift as in Definition 3. Let $\boldsymbol{\Phi} = \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top$ be the projection matrix onto the subspace spanned by the first k principal components of the training data. Then, the expected risk of PCR under covariate shift is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] &= (\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta}) \\ &+ \frac{\sigma_\nu^2}{n}\operatorname{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}_S^{-1}\hat{\boldsymbol{V}}\right) + \sigma_T^2,\end{aligned} \tag{9}$$

with $\boldsymbol{\beta}_T = r_w \boldsymbol{W}\boldsymbol{C}_{z,T}(\boldsymbol{I}_d + r_w^2\boldsymbol{C}_{z,T})^{-1}\boldsymbol{\theta}$ and $\sigma_T^2 = \sigma_\varepsilon^2 + \boldsymbol{\theta}^\top(\boldsymbol{I}_d + r_w^2\boldsymbol{C}_{z,T})^{-1}\boldsymbol{C}_{z,T}\boldsymbol{\theta}$.

This result resembles (4) but also shows that the introduction of covariate shift complicates the analysis because we have to deal with quantities from training and test distribution. Therefore, when inspecting the asymptotic result as stated below, we have to treat factors in the squared bias term individually according to the contribution from training and test data.

Covariate-shifted asymptotic PCR risk. To obtain asymptotic expressions of (9), we define the *asymptotic squared bias* term as

$$\begin{aligned}\operatorname{Bias}_{\gamma,T}(\hat{\boldsymbol{\theta}})^2 = &\begin{bmatrix}\bar{\boldsymbol{\beta}}_T^\top & \bar{\boldsymbol{\beta}}^\top\end{bmatrix} \\ &\begin{bmatrix}\boldsymbol{\Lambda}_{d,T} & -\boldsymbol{\Lambda}_{d,T}\boldsymbol{P}_k \\ -\boldsymbol{P}_k\boldsymbol{\Lambda}_{d,T} & \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T}\boldsymbol{P}_k\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{\beta}}_T \\ \bar{\boldsymbol{\beta}}\end{bmatrix},\end{aligned} \tag{10}$$

with $\bar{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}}_T = \boldsymbol{W}^{-1}\boldsymbol{\beta}_T$ and $\boldsymbol{\Lambda}_{d,T} \in \mathbb{R}^{d\times d}$ as the truncation of $\boldsymbol{\Lambda}_T$ to the first d dimensions. The *asymptotic variance* term is

$$\begin{aligned}\operatorname{Var}_{\gamma,T}(\hat{\boldsymbol{\theta}}) = &\frac{\sigma_\nu^2}{n}\Bigg(\operatorname{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T} + \boldsymbol{I}_k)\frac{1}{\mu(\boldsymbol{\Lambda},\gamma)}\right] \\ &+ (p-d)\int_{s_c}^{(1+\sqrt{\gamma})^2}\frac{1}{s}dF_\gamma(s)\Bigg),\end{aligned} \tag{11}$$

with $\mu(\boldsymbol{\Lambda},\gamma)$, F_γ and s_c as described in the non-covariate shifted variance term (7). Combining both results yields the following theorem for the asymptotic risk of PCR under covariate shift.

Theorem 4 (Covariate-shifted asymptotic PCR risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0,\infty)$, the expected risk of PCR under covariate shift will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] \to \operatorname{Bias}_{\gamma,T}(\hat{\boldsymbol{\theta}})^2 + \operatorname{Var}_{\gamma,T}(\hat{\boldsymbol{\theta}}) + \sigma_T^2.$$

This theorem has the same implications as Theorem 2 where there is no covariate shift about choosing the number of principal components k and for regularizing the interpolation peak. The main difference for covariate shifts lies in the term $\boldsymbol{C}_{z,T}$ for the squared bias and variance terms and in $\boldsymbol{C}_{z,S}$ for the estimation of $\hat{\boldsymbol{\theta}}$. Since their eigenvectors $\boldsymbol{V}$ are considered to be equal, we only have to consider differing eigenvectors $\boldsymbol{\Lambda}_S$, $\boldsymbol{\Lambda}_T$ which are diagonal matrices. Thus the shift between training and test covariance determines the risk.

4.2. Risk of baseline methods

Similar to Section 3.2, the results for the full linear regression as baseline model are special cases of the main PCR results with $k = p$. However, in contrast to the case without covariate shift, here the squared bias term will not diminish in the asymptotic case.

Let us start with the null predictor as a reference. For $\hat{\boldsymbol{\theta}} = 0$ the expected null risk under covariate shift becomes $\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] = \boldsymbol{\beta}_T^\top \boldsymbol{C}_T \boldsymbol{\beta}_T + \sigma_T^2$. As a generalization of the expected risk of full regression from (8), we obtain under covariate shift

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\beta}})\right] =&(\boldsymbol{\beta}_T - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\beta}) \\ &+ \frac{\sigma_\nu^2}{n}\operatorname{Tr}(\boldsymbol{C}_T\hat{\boldsymbol{C}}^{-1}) + \sigma_T^2.\end{aligned} \tag{12}$$

Finally, we can extend the asymptotic results from Theorem 3 to obtain the following result under covariate shift.

Theorem 5 (Covariate-shifted asymptotic full regression risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0,\infty)$, the expected risk of the full regression model under covariate shift will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\beta}})\right] \to \operatorname{Bias}_{\gamma,T}(\hat{\boldsymbol{\beta}})^2 + \operatorname{Var}_{\gamma,T}(\hat{\boldsymbol{\beta}}) + \sigma_T^2$$

with the asymptotic squared bias term as

$$\operatorname{Bias}_{\gamma,T}(\hat{\boldsymbol{\beta}})^2 = (\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}})^\top \boldsymbol{\Lambda}_{d,T}(\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}})$$

for $\gamma < 1$ and as in Theorem 4 for $\gamma \geq 1$ with $k = p$. The variance term $Var_\gamma(\hat{\boldsymbol{\beta}})$ is equal to the definition of the variance in Theorem 4 with $k = p$.

5. Numerical results

5.1. Simulation of in-distribution data

Setup. We focus on the practically relevant scenario of high-dimensional regression with a fixed relation between the number of features p and samples n controlled by γ. When varying γ we keep the number of spike eigenvalues d constant to simulate regression problems of varying size with fixed latent space dimension. While we select one set of parameters for visualizing the results, we have rigorously validated our results across numerous combinations. For the data-generating process, we choose the parameters $\alpha = 0.1$, $\sigma_\varepsilon = 0.1$, $r_\theta = 1$ and $\rho_x = 1$. We choose $d = 10$ spikes and vary k to see the effect of model misspecification. For our simulations, we choose $n = 500$ and set p accordingly to fulfill $\gamma = \frac{p}{n}$. We vary $\gamma \in [0.1, 30]$, i.e. from low-dimensional $\gamma < 1$ to high-dimensional $\gamma > 1$. We compute the risk $\mathbb{E}_\nu [R(\boldsymbol{\theta})]$ and present median values of the simulation results from 50 realizations as well as 25%, 75% quantiles. We provide code to reproduce the numerical simulations https://github.com/dgedon/PCR_spiked_covariance.

Main result. The main result for the risk analysis is presented in Figure 5. We can observe the following points: (1) Theory and simulation align well and therefore support our analysis. (2) For misspecified models with $k < d$ we remove important eigendirections in the PCA step and therefore the risk rises heavily; misspecifications with $k \gg d$ diminish the regularizing effect of PCR by including many noise directions and therefore get close to the full regression solution. (3) For appropriately[2] chosen k PCR does not suffer from the interpolation peak and therefore shows a desirable regularizing effect. (4) For $k \geq d$, the PCR model matches the full regression solution in the limit of small and large γ.

Bias-variance decomposition. We split the risk according to Theorem 2 and 3 in bias and variance terms to analyse them independently. The results are shown in Figure 6 where we can observe that the bias term is dominant for $k < d$. According to Theorem 2 for $k < d$ the least amount of risk is subtracted because $\boldsymbol{P}_k$ contains many zeros. For appropriately chosen k, the bias decreases with large γ. The variance term increases with larger k as supported by Theorem 2 because more noise directions are considered.

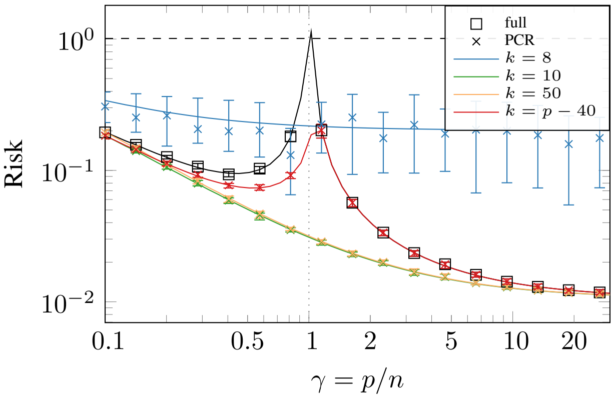


Figure 5: **Risk on in-distribution data: Simulation vs. analysis.** The different marks denote finite sample risk; solid lines denote analytical results. Null risk is given by the dashed line.

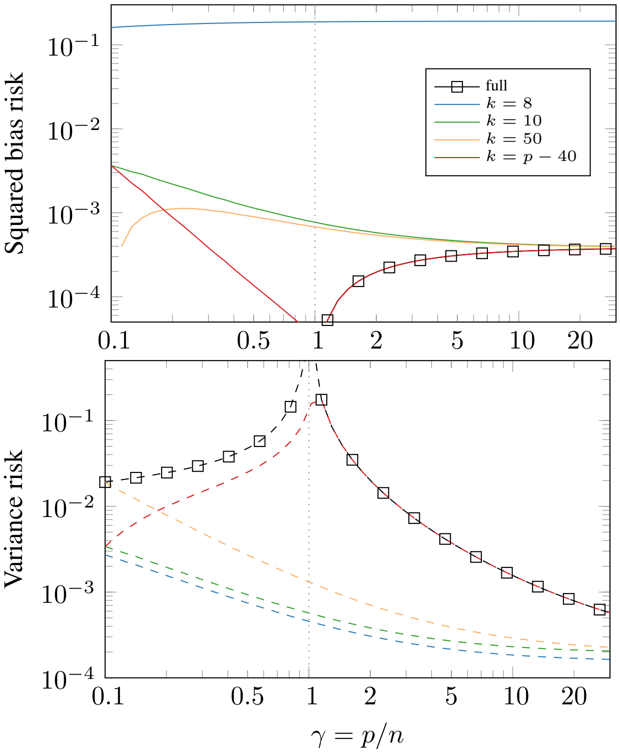


Figure 6: **Bias-variance decomposition of in-distribution risk: analysis.** Squared bias (top) and variance (bottom).

PCA-projection space. The linear regression within PCR operates through the projection of PCA onto the principal components with $k \leq p$ on a different space than the full regression from all features p. In this experiment, in contrast to varying γ with fixed k, we keep γ fixed and vary the ratio $\kappa = k/n$. Therefore, κ represents the effective dimension the linear regression operates on. In Figure 7 the result for varying κ is given. We can see that for $\kappa \to 1$, the PCR risk approaches the risk of full regression. Note that for $k > \min(p, n)$ there are no principal components left and the risk plateaus. This is also the reason why $\kappa \in [0, 1]$.

[2] I.e. $k = d + \Delta$ with Δ a small non-negative integer such that a limited number of noise directions are considered.

Curth et al. (2024) describes a similar setting where $\kappa > 1$ is reached by excess features that "only contribute to the creation of a richer basis".

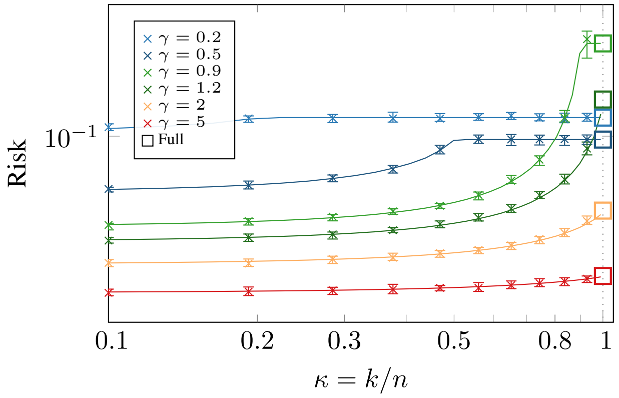


Figure 7: **Risk on PCA projected space: Simulation vs. analysis.** Varying $\kappa k/n$ with fixed γ.

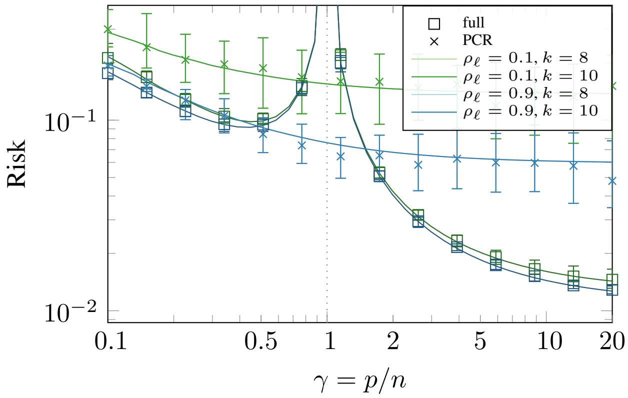


Figure 8: **Risk on covariate-shifted data: Simulation vs. analysis.** Different marks denote finite sample risk; solid lines denote analytical results. In green tones are results for low-correlated features $\rho_\ell = 0.1$; and in blue tones for high-correlated features $\rho_\ell = 0.9$.

5.2. Simulation of covariate-shifted data

Setup. We choose the same setup as for in-distribution data with $\alpha = 1$. The spike eigenvalues for training and test data are sampled i.i.d. from a zero-mean Normal distribution with $\mathrm{Cov}(u_S, u_T) = \sigma_\ell^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ such that $\lambda_{S,i} = \exp(\alpha u_{S,i})$, $\lambda_{T,i} = \exp(\alpha u_{T,i})$. By choosing σ_ℓ^2, ρ, we can control the covariate shift between the training and test data. Specifically, there are two scenarios without covariate shifts. (1) with $\sigma_\ell^2 = 0$, we have $\lambda_{i,S} = \lambda_{i,T} = 1$. (2) with $\sigma_\ell^2 > 0$ but $\rho = 1$, we have $\lambda_{i,S} = \lambda_{i,T} \neq 1$. Finally, we can create another scenario (3) where we introduce covariate shift with $\sigma_\ell^2 > 0$ and $\rho \in]0, 1[$ which defines correlated but different latent factors. Thus, σ_ℓ^2 and ρ control the correlation between training and test data. We focus on the covariate shift by choosing $\sigma_\ell^2 = 1$ and investigate the effect of correlation between training and test data through different choices of ρ_ℓ.

Main result. In Figure 8 we present results for low correlated covariate eigenvalues with $\rho_\ell = 0.1$ and highly correlated eigenvalues with $\rho_\ell = 0.9$. We notice both scenarios have qualitatively similar behaviour which follows the observations from the in-distribution results in Figure 5. The main difference is that highly correlating train and test data leads to lower risk. Since test data in this scenario is generated roughly from the same distribution as the training data, the model is less misspecified which leads to lower risk.

6. Related work

Overparameterization. The 'double descent' phenomenon (Belkin et al., 2019) for the generalization curve of overparameterized models was already discovered and analysed in early works (Krogh & Hertz, 1991; Geman et al., 1992; Opper, 1995). While it has been observed in deep state-of-the-art models (d'Ascoli et al., 2020; Nakkiran et al., 2021), most theoretical studies focus on simple models. Examples are found for linear regression (Bartlett et al., 2020; Muthukumar et al., 2020; Hastie et al., 2022), ensembles (LeJeune et al., 2020; Loureiro et al., 2021), classification (Gerace et al., 2020; Wang et al., 2021; Deng et al., 2022), random features (Belkin et al., 2019; Mei & Montanari, 2022) or small neural networks trained using gradient descent (Advani et al., 2020; Frei et al., 2022; Cao et al., 2022). Practical model regularizers such as Ridge are analysed in Tsigler & Bartlett (2023). PCR falls widely into the category of sparsity-inducing regularizers. For Lasso regularization bounds for ℓ_1-norm interpolators are provided (Chatterji & Long, 2022; Wang et al., 2022) but also arbitrary norms are studied (Koehler et al., 2021). We extend the analysis of sparsity-inducing regularizers to the commonly used PCR model. We further broaden our analysis to distribution shifts, a relatively underexplored area in the context of overparameterization (Tripuraneni et al., 2021a;b; Emami et al., 2020).

PCR analysis. Using PCA (Pearson, 1901; Jolliffe, 2002) is common—discussions focus on the choice of principle components (Breiman & Freedman, 1983) or high-dimensional data (Lee et al., 2012). Since PCA acts on the eigenvectors of the covariance matrix, it can be viewed as a spectral regularizer. Finite sample risk bounds were analysed for general spectral regularization including PCR (Bauer et al., 2007; Dicker et al., 2017) and adaptive PCR under the same latent factor model as ours but using concentration inequalities (Bing et al., 2021). Hucker &

Wahl (2023) provide high probability bounds for the PCR risk. Furthermore, PCR is investigated in Xu & Hsu (2019) for general but fully known covariances $\boldsymbol{C}$ in the asymptotic regime.Wu & Xu (2020) extend it by showing that the misalignment of true and estimated eigenvectors affects the risk. Huang et al. (2022) use misalignment bounds (Loukas, 2017) to remove the known covariance assumption and obtain non-asymptotic risk bounds. We extend and complement existing analyses to the case real-world sample covariances $\hat{\boldsymbol{C}}$ to obtain asymptotic risk guarantees for the latent factor model using random matrix theory.

Spiked covariance model. The spiked covariance model was introduced in Johnstone (2001) for high-dimensional covariance matrices where data is generated from a low-dimensional subspace. It is a popular model for covariance estimation (Donoho et al., 2018; Dobriban et al., 2020) and has been used to analyse the performance of PCA (Johnstone & Paul, 2018; Bai & Silverstein, 2010). We combine the spiked covariance model with the latent factor model to obtain a regression model with low-dimensional latent factors and high-dimensional covariates. Then, we apply asymptotic results from random matrix theory for the eigenvalue shift of the spikes (Baik et al., 2005; Baik & Silverstein, 2006), and the eigenvector misalignment (Paul, 2007; Johnstone et al., 2009) to obtain asymptotic risk guarantees for PCR.

7. Conclusion

Guide to choosing k. Deriving precise guidelines from our analysis is challenging. From Theorems 2 and 4, we can see that choosing a larger number of principal components k increases the variance due to noise contributions. From Figures 5 and 8, we deduce that choosing k too small increases the risk due to discarding signal components. In practice, for $\gamma = \frac{p}{n}$ small < 0.5 or large > 2, k has little effect on the risk.

Summary. We present asymptotic results for the generalization risk of PCR under the spiked covariance model based on random matrix theory. Furthermore, we consider the case where the training and test distribution vary. Our analysis generalizes the asymptotic result for linear regression from Hastie et al. (2022) which is a special case of PCR without dimensionality reduction ($k = p$). In the non-asymptotic regime, Huang et al. (2022) show similar results, thus independently supporting our findings. Selecting the correct number of principal components k is crucial for the risk as Theorems 2 and 4 suggest.

While our results that PCA mitigates the interpolation peak due to its regularizing behaviour may not be a surprise, we provide formal guarantees for the risk of a commonly used model on real-world data structures. Practitioners can now rely on fundamental guarantees for model development, but more research is needed for general data structures.

Limitations and future work. In this paper, we limit the theoretical analysis to the supervised case where data is sampled from the spiked covariance model with linear regressors. However, the risk on the MAGIC wheat genetics data set in Figure 1 qualitatively resemble our numerical results in Figure 5 which suggests that our results can closely replicate results that are not linearly separable. Our analysis is based on random matrix theory and therefore only holds in the asymptotic regime. For finite sample risk bounds, we refer to Bing et al. (2021).

Extending our analysis to more general covariance matrices $\boldsymbol{C}$ is an important extension and the tools we developed could be exploited since the Stieltjes transformation holds in more general settings. We believe the phenomenon we study with decaying eigenvalues is the most relevant for generalization analysis. This is extensively studied e.g. in Bartlett et al. (2020). One way to include the effect of gradient-based training could be to generalize our findings to spectral regularization techniques such as Landweber iteration (Bauer et al., 2007). Another interesting avenue is the extension into semi-supervised settings (Wasserman & Lafferty, 2007). Scenarios, where the PCA is trained on a large unlabeled dataset and then used for regression on a small labeled dataset, are related to the common idea of pre-training large models and might shed light on the generalization behaviour of such models.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation an by the Kjell och Märta Beijer Foundation. We thank Parthe Pandit for essential discussions about the covariate shift part of the paper.

Impact statement

This paper presents work aimed at advancing the field of machine learning by providing fundamental insights into the generalization properties of PCR under data from the spiked covariance model. While we primarily contribute to theoretical understanding and empirical validation of our results, there are practical insights for practitioners of the widely adopted PCR. Namely, now there exist fundamental generalisation guarantees for real-world data structures in high-dimensional settings and practitioners can design models with the knowledge that there is no interpolation peak.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Anderson, G. W., Guionnet, A., and Zeitouni, O. *An introduction to random matrices*. Cambridge university press, 2010.

Bach, F. High-dimensional analysis of double descent for linear regression with random projections. *arXiv preprint arXiv:2303.01372*, 2023.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Baik, J. and Silverstein, J. W. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

Baik, J., Arous, G. B., and Péché, S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bauer, F., Pereverzev, S., and Rosasco, L. On regularization algorithms in learning theory. *Journal of complexity*, 23 (1):52–72, 2007.

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Belkin, M., Hsu, D. J., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. Prediction under latent factor regression: Adaptive PCR, interpolating predictors and beyond. *The Journal of Machine Learning Research*, 22(1):7994–8043, 2021.

Bloemendal, A. and Virág, B. Limits of spiked random matrices i. *Probability Theory and Related Fields*, 156: 795–825, 2013.

Breiman, L. and Freedman, D. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35: 25237–25250, 2022.

Chatterji, N. S. and Long, P. M. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022.

Curth, A., Jeffares, A., and van der Schaar, M. A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Dicker, L. H., Foster, D. P., and Hsu, D. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022 – 1047, 2017.

Dobriban, E., Leeb, W., and Singer, A. Optimal prediction in the linearly transformed spiked model. *The Annals of Statistics*, 48(1):491 – 513, 2020.

Donoho, D. L., Gavish, M., and Johnstone, I. M. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pp. 2892–2901. PMLR, 2020.

Frei, S., Chatterji, N. S., and Bartlett, P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Geweke, J. Bayesian reduced rank regression in econometrics. *Journal of econometrics*, 75(1):121–146, 1996.

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

Huang, N. T., Hogg, D. W., and Villar, S. Dimensionality reduction, regularization, and generalization in overparameterized regressions. *SIAM Journal on Mathematics of Data Science*, 4(1):126–152, 2022.

Hucker, L. and Wahl, M. A note on the prediction error of principal component regression in high dimensions. *Theory of Probability and Mathematical Statistics*, 109: 37–53, 2023.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 29(2):295–327, 2001.

Johnstone, I. M. and Paul, D. PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

Johnstone, I. M., Lu, A. Y., Nadler, B., Witten, D. M., Hastie, T., Tibshirani, R., and Ramsay, J. O. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–703, 2009.

Jolliffe, I. T. *Principal component analysis for special types of data*. Springer, 2002.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Krogh, A. and Hertz, J. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Lee, Y. K., Lee, E. R., and Park, B. U. Principal component analysis in very high-dimensional spaces. *Statistica Sinica*, pp. 933–956, 2012.

LeJeune, D., Javadi, H., and Baraniuk, R. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 3525–3535. PMLR, 2020.

Loukas, A. How close are the eigenvectors of the sample and actual covariance matrices? In *International Conference on Machine Learning*, pp. 2228–2237. PMLR, 2017.

Loureiro, B., Sicuro, G., Gerbelot, C., Pacco, A., Krzakala, F., and Zdeborová, L. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

Marčenko, V. A. and Pastur, L. A. The distribution of eigenvalues in certain sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Massy, W. F. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Opper, M. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pp. 922–925, 1995.

Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Scott, M. F., Fradgley, N., Bentley, A. R., Brabbs, T., Corke, F., Gardner, K. A., Horsnell, R., Howell, P., Ladejobi, O., Mackay, I. J., et al. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome biology*, 22(1):1–30, 2021.

Stieltjes, T.-J. Recherches sur les fractions continues. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 8, pp. J1–J122, 1894.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2023.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Tripuraneni, N., Adlam, B., and Pennington, J. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021a.

Tripuraneni, N., Adlam, B., and Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34:13883–13897, 2021b.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

Vijayakumar, S. and Schaal, S. Locally weighted projection regression: An o(n) algorithm for incremental real time learning in high dimensional space. In *Proceedings of the seventeenth international conference on machine learning (ICML 2000)*, volume 1, pp. 288–293, 2000.

Wang, G., Donhauser, K., and Yang, F. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10572–10602. PMLR, 2022.

Wang, K. and Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.*, 32(2):108–118, 2008.

Wang, K., Muthukumar, V., and Thrampoulidis, C. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021.

Wasserman, L. and Lafferty, J. Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, 20, 2007.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xu, J. and Hsu, D. J. On the number of variables to use in principal component regression. *Advances in neural information processing systems*, 32, 2019.

Yang, J. and Johnstone, I. M. Edgeworth correction for the largest eigenvalue in a spiked pca model. *Statistica Sinica*, 28(4):2541, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A. Experiment details: genetics example

Background. The Diverse MAGIC Wheat data set[3] is based on 16 founding wheat varieties which were listed between 1935 and 2004. These varieties were interbred to obtain new wheat varieties. From the resulting wheat types, the genome of a total of 502 kinds of wheat were sequenced. This genome sequence consists of approximately 1.1 million single nucleotide polymorphisms. Furthermore, phenotypes of the 502 wheat types were analysed, see Scott et al. (2021).

Data processing. The genotypes consist of binary features. The binary variables represent equality or difference to a reference genotype. The phenotypes are real-values variables. We choose the phenotype column named 'HET_2' in this example. Missing values for both, genotype and phenotype are replaced with the mean value of the variable. We select a subset of genotypes as inputs randomly at uniform to obtain the necessary p features. Then, we normalize both, genotype and phenotype by z-transformation.

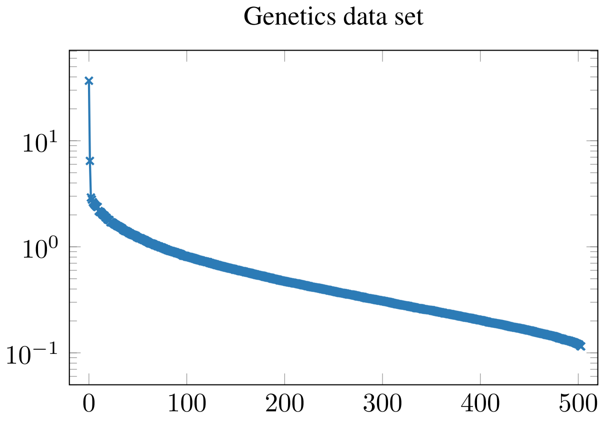


Figure A-1: **Eigenvalue distribution of the Diverse MAGIC Wheat genetics data set.**

Data analysis. In Figure A-1 we plot the eigenvalue distribution for the Diverse MAGIC Wheat data set. We observe that the eigenvalue distribution is heavy-tailed. While it has two dominant eigenvalues, it does not depict a clear example of a low-dimensional latent manifold. Therefore, using the PCR model will discard some useful information. We also note, that there is no reason to assume a linear relationship between the features and the outcome which is in contrast to our synthetically generated data.

[3] http://mtweb.cs.ucl.ac.uk/mus/www/MAGICdiverse/

B. Additional numerical results

B.1. Ridge regression

We compare the PCR model with different values for Ridge regression for in-distribution data. The results are depicted in Figure B-2. Here, we compare with the optimal $k = d$ PCR model. Note that solid lines indicate analytical solutions while markers indicate simulations. Analytical solutions for Ridge regression could be obtained, see Hastie et al. (2022). We can observe that Ridge regression shows similar regularizing behaviour as regularization through PCR. However, the exact shape of the risk curve varies and for the case we present, PCR with $k = d$ shows the lowest risk for all γ.

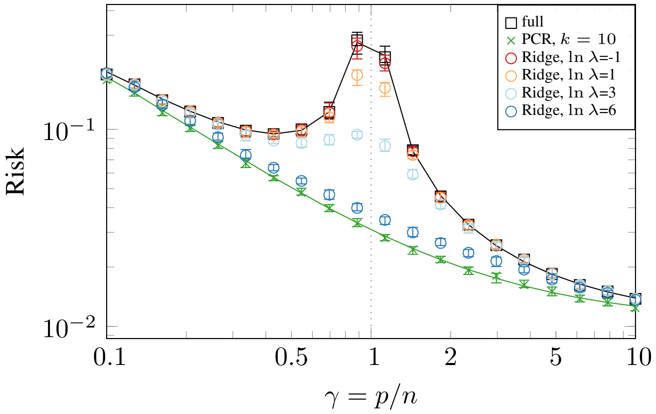


Figure B-2: **Risk PCR vs. Ridge regularisation.** We show median, 25%, and 75% quantiles over 50 seeds for all simulation results.

B.2. Varying output noise σ_ν

In Figure B-3, we compare the PCR model under varying output noise σ_ν for in-distribution data. Results from Theorem 2 are compared with simulations. The figure shows that our analysis holds for a suitable range of noise values and that the risk decreases with lower noise variance.

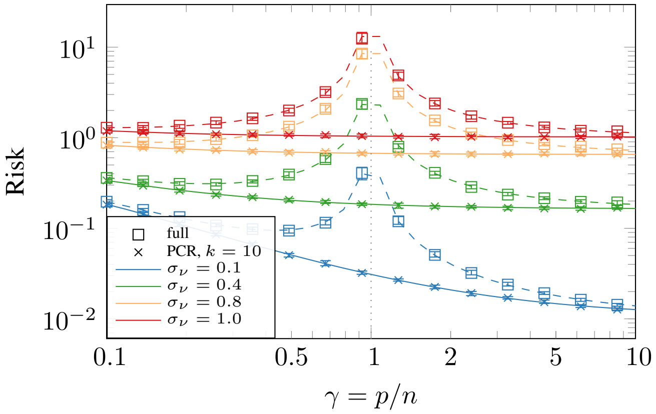


Figure B-3: **Risk PCR for varying σ_ν: Simulation vs. analysis.** We show median, 25%, and 75% quantiles over 50 seeds for all simulation results. Dashed lines are analytical solutions for the full regression; solid lines are analytical solutions for the PCR.

C. Proofs

In this section, we will follow the structure of the main paper to prove all results. For clarity, we will repeat the results before providing the proof.

C.1. Data generator

In this section, we derive the linear model from the latent factor model in Definition 1. Let us first restate the definition:

Definition 1. The *latent factor model* is the linear model $\boldsymbol{x}_i = \boldsymbol{W} r_w \boldsymbol{z}_i + \boldsymbol{e}_i$, and $y_i = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \varepsilon_i$. With latent variables $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_z)$, with diagonal covariance $\boldsymbol{C}_z$, feature noise $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, outcome noise $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, feature matrix $\boldsymbol{W} \in \mathbb{R}^{p \times d}$ such that $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_d$. Let $r_w^2 = \frac{p}{\mathrm{Tr}(\boldsymbol{\Lambda}_z)} \rho_x$ to control the feature signal-to-noise-ratio (SNR) $\rho_x = \frac{\mathbb{E}\left[\|\boldsymbol{W} r_w \boldsymbol{z}\|_2^2\right]}{\mathbb{E}\left[\|\boldsymbol{e}\|_2^2\right]}$, label noise $\varepsilon_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_\varepsilon^2)$ and let $\boldsymbol{\theta} = 1_d \frac{r_\theta}{\sqrt{\mathrm{Tr}(\boldsymbol{\Lambda}_z)}}$ to control the outcome SNR $\rho_y = \frac{\mathbb{E}\left[\|\boldsymbol{\theta}^\top \boldsymbol{z}\|_2^2\right]}{\mathbb{E}\left[\|\varepsilon\|_2^2\right]} = \frac{r_\theta^2}{\sigma_\varepsilon^2}$.

We aim to derive the linear model $y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \nu_i$, i.e. to find the expressions for

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}), \tag{C-1}$$

$$\nu_i \sim \mathcal{N}(0, \sigma_\nu^2) \quad \text{with} \quad \sigma_\nu^2 = \sigma_\varepsilon^2 + \boldsymbol{\theta}^\top (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{C}_z \boldsymbol{\theta}, \tag{C-2}$$

$$\boldsymbol{\beta} = r_w \boldsymbol{W} \boldsymbol{C}_z (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{\theta}. \tag{C-3}$$

Proof. The covariance matrix of $(\boldsymbol{x}_i, y_i)$ under the linear model is given by

$$\begin{bmatrix} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^\top\right] & \mathbb{E}\left[y_i \boldsymbol{x}_i^\top\right] \\ \mathbb{E}\left[y_i^\top \boldsymbol{x}_i\right] & \mathbb{E}\left[y_i y_i^\top\right] \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{\beta}^\top \boldsymbol{C} \\ \boldsymbol{C}\boldsymbol{\beta} & \boldsymbol{\beta}^\top \boldsymbol{C} \boldsymbol{\beta} + \sigma_\nu^2 \end{bmatrix}. \tag{C-4}$$

We can compare this with the covariance under the latent factor model

$$\begin{bmatrix} \mathbb{E}\left[\boldsymbol{W} \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{W}^\top r_w^2 + \boldsymbol{e}_i \boldsymbol{e}_i^\top\right] & \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{W}^\top r_w + \varepsilon_i \boldsymbol{e}_i^\top\right] \\ \mathbb{E}\left[r_w \boldsymbol{W} \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{\theta} + \boldsymbol{e}_i \varepsilon_i^\top\right] & \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{z}_i \boldsymbol{z}_i \boldsymbol{\theta} + \varepsilon_i \varepsilon_i^\top\right] \end{bmatrix} = \begin{bmatrix} \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{W}^\top r_w^2 + \boldsymbol{I}_p & \boldsymbol{\theta}^\top \boldsymbol{C}_z \boldsymbol{W}^\top r_w \\ r_w \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{\theta} & \boldsymbol{\theta}^\top \boldsymbol{C}_z \boldsymbol{\theta} + \sigma_\varepsilon^2 \end{bmatrix}. \tag{C-5}$$

Comparing element $(1,1)$, we directly observe that $\boldsymbol{C} = \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{W}^\top r_w^2 + \boldsymbol{I}_p$ which is our spiked covariance model.

From element $(2,1)$, we get the following when using $\boldsymbol{C}$ as well as $\boldsymbol{W} \in \mathbb{R}^{p \times d}$ such that $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_d$

$$\boldsymbol{C}\boldsymbol{\beta} = r_w \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{\theta} \tag{C-6}$$

$$\boldsymbol{\beta} = \begin{bmatrix} (\boldsymbol{C}_z r_w^2 + \boldsymbol{I}_d)^{-1} & 0 \\ 0 & \boldsymbol{I}_{p-d} \end{bmatrix} r_w \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{\theta} \tag{C-7}$$

$$= (\boldsymbol{I}_p + \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{W}^\top r_w^2)^{-1} r_w \boldsymbol{W} \boldsymbol{C}_z \boldsymbol{\theta}. \tag{C-8}$$

Using the push-through or Woodbury matrix identity, we obtain the result for $\boldsymbol{\beta}$.

Finally from element $(2,2)$, we get the following expression

$$\sigma_\nu^2 = \boldsymbol{\theta}^\top \boldsymbol{C}_z \boldsymbol{\theta} - \boldsymbol{\beta}^\top \boldsymbol{C} \boldsymbol{\beta} + \sigma_\varepsilon^2. \tag{C-9}$$

Using (C-8) for $\boldsymbol{\beta}^\top$ with the result for $\boldsymbol{\beta}$ and $\boldsymbol{C}$, we obtain

$$\sigma_\nu^2 = \boldsymbol{\theta}^\top (\boldsymbol{C}_z - \boldsymbol{C}_z \boldsymbol{C}_z r_w^2 (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1}) \boldsymbol{\theta} + \sigma_\varepsilon^2 \tag{C-10}$$

$$= \boldsymbol{\theta}^\top \boldsymbol{C}_z (\boldsymbol{I}_d - \boldsymbol{C}_z r_w^2 (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1}) \boldsymbol{\theta} + \sigma_\varepsilon^2 \tag{C-11}$$

Using the identity $(\boldsymbol{I} + \boldsymbol{P})^{-1} = \boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{P})^{-1} \boldsymbol{P}$, we obtain

$$\sigma_\nu^2 = \sigma_\varepsilon^2 + \boldsymbol{\theta}^\top \boldsymbol{C}_z (\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1} \boldsymbol{\theta}. \tag{C-12}$$

Using the push-through or Woodbury identity yields the result for the variance. □

C.2. Analysis of risk

C.2.1. Risk of PCR

PCR parameter estimator. Let us first re-state the parameter estimator for PCR:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + \hat{\boldsymbol{S}}^{-1} \hat{\boldsymbol{U}}^\top \boldsymbol{\nu}. \tag{3}$$

Proof. We consider the unregularized linear regression solution between the latent variables $\hat{\boldsymbol{Z}}$ and the outcome $\boldsymbol{y}$:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}})^+ \hat{\boldsymbol{Z}}^\top \boldsymbol{y} && \text{(C-13)} \\ &= (\hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} \hat{\boldsymbol{S}}_k)^+ \hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{U}}^\top \boldsymbol{y} && \text{(C-14)} \end{aligned}$$

with $\hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} = \boldsymbol{I}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\nu}$

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{S}}_k)^+ \hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{U}}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\nu}) && \text{(C-15)} \\ &= (\hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{S}}_k)^+ \hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{S}} \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + (\hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{S}}_k)^+ \hat{\boldsymbol{S}}_k^\top \hat{\boldsymbol{U}}^\top \boldsymbol{\nu} && \text{(C-16)} \end{aligned}$$

Where we used $\boldsymbol{X} = \hat{\boldsymbol{U}} \hat{\boldsymbol{S}} \hat{\boldsymbol{V}}^\top$. Now we combine the singular value matrices. We indicate the dimensions of combined matrices. Note that $\hat{\boldsymbol{S}}_k$ and $\hat{\boldsymbol{S}}$ are of different sizes.

$$\begin{aligned} \hat{\boldsymbol{\theta}} = &\begin{bmatrix} \frac{1}{\hat{\sigma}_1^2} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \frac{1}{\hat{\sigma}_k^2} \end{bmatrix}_{k \times k} \left[\begin{array}{ccc|c} \hat{\sigma}_1^2 & & \boldsymbol{0} & \\ & \ddots & & \boldsymbol{0} \\ \boldsymbol{0} & & \hat{\sigma}_k^2 & \end{array}\right]_{k \times p} \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} \quad + \\ &+ \left[\begin{array}{ccc|c} \frac{1}{\hat{\sigma}_1} & & \boldsymbol{0} & \\ & \ddots & & \boldsymbol{0} \\ \boldsymbol{0} & & \frac{1}{\hat{\sigma}_k} & \end{array}\right]_{k \times n} \hat{\boldsymbol{U}}^\top \boldsymbol{\nu} && \text{(C-17)} \\ = &\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \end{bmatrix} \hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + \begin{bmatrix} \hat{\boldsymbol{S}}_{kk}^{-1} & \boldsymbol{0} \end{bmatrix} \hat{\boldsymbol{U}}^\top \boldsymbol{\nu} && \text{(C-18)} \end{aligned}$$

Summarizing the matrices by truncating $\hat{\boldsymbol{V}}^\top$ and $\hat{\boldsymbol{U}}^\top$ yields the following solution for the regression parameter estimation

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{V}}_k^\top \boldsymbol{\beta} + \hat{\boldsymbol{S}}_{kk}^{-1} \hat{\boldsymbol{U}}_k^\top \boldsymbol{\nu}. \tag{C-19}$$

This concludes the proof as we defined $\hat{\boldsymbol{V}} := \hat{\boldsymbol{V}}_k$, $\hat{\boldsymbol{S}} := \hat{\boldsymbol{S}}_{kk}$ and $\hat{\boldsymbol{U}} := \hat{\boldsymbol{U}}_k$. □

Expected risk of PCR. Let us first re-state the expected risk result for PCR:

$$\mathbb{E}_{\boldsymbol{\nu}} \left[R(\hat{\boldsymbol{\theta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi} \boldsymbol{C} \boldsymbol{\Pi} \boldsymbol{\beta} + \frac{\sigma_\nu^2}{n} \operatorname{Tr} \left(\hat{\boldsymbol{V}}^\top \boldsymbol{C} \hat{\boldsymbol{V}} \hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1} \hat{\boldsymbol{V}} \right) + \sigma_\nu^2. \tag{4}$$

Proof. We define the risk as the expectation over the mean squared error and then use $y_0 = \boldsymbol{\beta}^\top \boldsymbol{x}_0 + \nu_0$, as well as $\hat{y}(\boldsymbol{x}_0) = \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}} = \hat{\boldsymbol{V}}^\top \boldsymbol{x}_0$ to obtain

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{(\boldsymbol{x}_0, y_0)} \left[(y_0 - \hat{y}(\boldsymbol{x}_0))^2 \right] && \text{(C-20)} \\ &= \mathbb{E}_{\boldsymbol{x}_0} \left[(\boldsymbol{\beta}^\top \boldsymbol{x}_0 + \nu_0 - \hat{y}(\boldsymbol{x}_0))^2 \right] && \text{(C-21)} \\ &= \mathbb{E}_{\boldsymbol{x}_0} \left[(\boldsymbol{\beta}^\top \boldsymbol{x}_0 + \nu_0 - \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{z}})^2 \right] && \text{(C-22)} \\ &= \mathbb{E}_{\boldsymbol{x}_0} \left[(\boldsymbol{\beta}^\top \boldsymbol{x}_0 + \nu_0 - \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{V}}^\top \boldsymbol{x}_0)^2 \right] && \text{(C-23)} \\ &= \mathbb{E}_{\boldsymbol{x}_0} \left[\left((\boldsymbol{\beta} - \hat{\boldsymbol{V}} \hat{\boldsymbol{\theta}})^\top \boldsymbol{x}_0 + \nu_0 \right)^2 \right] && \text{(C-24)} \\ &= (\boldsymbol{\beta} - \hat{\boldsymbol{V}} \hat{\boldsymbol{\theta}})^\top \boldsymbol{C} (\boldsymbol{\beta} - \hat{\boldsymbol{V}} \hat{\boldsymbol{\theta}}) + \nu_0 \nu_0^\top && \text{(C-25)} \end{aligned}$$

Define the orthogonal projector $\boldsymbol{\Phi} = \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top$ and define another orthogonal projector with $\boldsymbol{\Pi} = \boldsymbol{I}_p - \boldsymbol{\Phi}$. For simplicity, let us rephrase the following

$$\begin{aligned} \boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{\theta}} &= \boldsymbol{\beta} - \hat{\boldsymbol{V}}(\hat{\boldsymbol{V}}^\top \boldsymbol{\beta} + \hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu) && \text{(C-26)} \\ &= \boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu && \text{(C-27)} \\ &= (\boldsymbol{I}_p - \boldsymbol{\Phi})\boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu && \text{(C-28)} \\ &= \boldsymbol{\Pi}\boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu && \text{(C-29)} \end{aligned}$$

Now let us use this expression to take the expectation of the risk w.r.t. the noise. This yields

$$\mathbb{E}_\nu \left[R(\hat{\boldsymbol{\theta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta} + \mathbb{E}_\nu \left[\mathrm{Tr}(\nu^\top \hat{\boldsymbol{U}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu) \right] + \mathbb{E}_\nu \left[\nu\nu^\top \right] \tag{C-30}$$

Here we made use of the Trace since the expression is scalar. Hence, we can use the cyclic property of the Trace and pull the expectation inside

$$\mathbb{E}_\nu \left[R(\hat{\boldsymbol{\theta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta} + \mathrm{Tr}(\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \mathbb{E}_\nu \left[\nu\nu^\top \right] \hat{\boldsymbol{U}}\hat{\boldsymbol{S}}^{-1}) + \mathbb{E}_\nu \left[\nu\nu^\top \right] \tag{C-31}$$

with $\mathbb{E}_\nu \left[\nu\nu^\top \right] = \sigma_\nu^2$ and $\hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} = \boldsymbol{I}$

$$\mathbb{E}_\nu \left[R(\hat{\boldsymbol{\theta}}) \right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta} + \sigma_\nu^2 \, \mathrm{Tr}(\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-2}) + \sigma_\nu^2 \tag{C-32}$$

Using $\hat{\boldsymbol{S}}^{-2} = \frac{1}{n}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{+}\hat{\boldsymbol{V}}$ we obtain (4) which concludes the proof. □

Eigenvector shift product. Let us first re-state the result which generalizes the eigenvector shift to matrices:

$$\boldsymbol{P}_k = \begin{cases} \mathrm{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \dots, (\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k)^2, 0, \dots, 0 \right) & \text{for} \quad k < d, \\ \mathrm{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \dots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2 \right) & \text{for} \quad k = d, \\ \mathrm{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2 + c_1^2, \dots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2 + c_d^2 \right) & \text{for} \quad k > d, \end{cases} \tag{5}$$

where $\boldsymbol{V}_{d,d} \in \mathbb{R}^{d \times d}$ and $\hat{\boldsymbol{V}}_d \in \mathbb{R}^{d \times k}$ are the matrices where the eigenvectors are truncated to the first d elements in each eigenvector; and with the correction factor $c_i^2 = (k-d)\frac{1-(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2}{p-d}$ for $k > d$.

Let us give an informal justification of (5): Define $\hat{\boldsymbol{V}}_d \in \mathbb{R}^{d \times k}$ and $\boldsymbol{V}_{d,d} \in \mathbb{R}^{d \times d}$ as the matrices where the eigenvectors are truncated to the first d eigenvector elements. Then, the expression $\boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} \in \mathbb{R}^{d \times d}$ describes the matrix equivalent of the eigenvector shift (2) which is defined only for $(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_j)^2$ with $i = j$. These are the diagonal terms of the matrix expression. For off-diagonal values with $i \neq j$ we have that $(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_j)^2 \to 0$ because as $p \to \infty$ the estimates eigenvectors $\hat{\boldsymbol{v}}_j$ are randomly placed in a p-dimensional space. Hence $P\left((\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_j)^2 > \epsilon \right) \to 0$ which means that the expression $\boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d}$ yields a diagonal matrix.

Dependent on the choice of k, there are three different results:

1. For $k < d$: since $\hat{\boldsymbol{V}}_d \in \mathbb{R}^{d \times k}$, we have that $\hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top$ will only have non-zero values for first k diagonal entries, yielding

$$\boldsymbol{P}_k = \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} = \mathrm{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \dots, (\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k)^2, 0, \dots, 0 \right). \tag{C-33}$$

2. For $k = d$: with the same reasoning as in the previous case we obtain

$$\boldsymbol{P}_k = \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} = \mathrm{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, \dots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2 \right). \tag{C-34}$$

3. For $k > d$: Let us rewrite the matrix product as

$$\begin{aligned} \boldsymbol{P}_k = \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} = &= \boldsymbol{V}_{d,d}^\top \begin{bmatrix} \hat{\boldsymbol{V}}_{d,:d} & \hat{\boldsymbol{V}}_{d,d:} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{V}}_{d,:d}^\top \\ \hat{\boldsymbol{V}}_{d,d:}^\top \end{bmatrix} \boldsymbol{V}_{d,d} && \text{(C-35)} \\ &= \boldsymbol{V}_{d,d}^\top \left(\hat{\boldsymbol{V}}_{d,:d} \hat{\boldsymbol{V}}_{d,:d}^\top + \hat{\boldsymbol{V}}_{d,d:} \hat{\boldsymbol{V}}_{d,d:}^\top \right) \boldsymbol{V}_{d,d} && \text{(C-36)} \\ &= \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_{d,:d} \hat{\boldsymbol{V}}_{d,:d}^\top \boldsymbol{V}_{d,d} + \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_{d,d:} \hat{\boldsymbol{V}}_{d,d:}^\top \boldsymbol{V}_{d,d} && \text{(C-37)} \end{aligned}$$

where the first term if equal to $\boldsymbol{P}_k$ for $k = d$. Let us write the second term element-wise as $\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_{i,d:} \hat{\boldsymbol{v}}_{i,d:}^\top \boldsymbol{v}_i$ which is equal to $\hat{\boldsymbol{v}}_{i,d:} \hat{\boldsymbol{v}}_{i,d:}^\top$ since $\boldsymbol{v}_{ij} = 0$ if $i \neq j$. Hence, we need to know the expected value of the elements $\hat{\boldsymbol{v}}_{ij}$ for $j > d$. To identify this, we can note that $1 = \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{v}}_i$ which we can expand as $1 = \sum_{j=1}^d \hat{\boldsymbol{v}}_{ij}^\top \hat{\boldsymbol{v}}_{ij} + \sum_{j=d+1}^p \hat{\boldsymbol{v}}_{ij}^\top \hat{\boldsymbol{v}}_{ij}$. Again, we can expand the first sum with $\boldsymbol{v}$ as $\boldsymbol{v}_{ij} = 1$ only for $i = j$. Hence, we obtain

$$1 = (\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2 + \sum_{j=d+1}^{p} \hat{\boldsymbol{v}}_{ij}^\top \hat{\boldsymbol{v}}_{ij} \tag{C-38}$$

Hence, assuming that $\boldsymbol{v}_{ij}$ are uniformly distributed, we get that $\hat{\boldsymbol{v}}_{ij}^\top \hat{\boldsymbol{v}}_{ij} = \frac{1-(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2}{p-d}$. Finally, we see that we have to sum $k - d$ of these elements in the expression which leads to the second term in (C-37) yielding

$$c_i^2 = (k-d)\frac{1-(\boldsymbol{v}_i^\top \hat{\boldsymbol{v}}_i)^2}{p-d}. \tag{C-39}$$

Finally, we obtain in the case for $k > d$

$$\boldsymbol{P}_k = \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} = \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2 + c_1^2, \ldots, (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2 + c_d^2\right). \tag{C-40}$$

Asymptotic PCR risk. In order to prove Theorem 2, we re-state the results for the bias-squared and the variance terms first: The *asymptotic bias-squared* term is given by

$$\text{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 = \bar{\boldsymbol{\beta}}^\top \left(\boldsymbol{\Lambda}_d - \boldsymbol{\Lambda}_d \boldsymbol{P}_k - \boldsymbol{P}_k \boldsymbol{\Lambda}_d + \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_z \boldsymbol{P}_k\right) \bar{\boldsymbol{\beta}} \tag{6}$$

with $\bar{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta} = r_w \boldsymbol{C}_z(\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1}\boldsymbol{\theta}$, and $\boldsymbol{\Lambda}_d \in \mathbb{R}^{d\times d}$ as the truncation of the population eigenvalue matrix $\boldsymbol{\Lambda}$ to the first d dimensions. The *asymptotic variance* term is

$$\text{Var}_\gamma(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n}\left(\operatorname{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k)\frac{1}{\mu(\boldsymbol{\Lambda},\gamma)}\right] + (p-d)\int_{s_c}^{(1+\sqrt{\gamma})^2} \frac{1}{s} dF_\gamma(s)\right) \tag{7}$$

with $\mu(\boldsymbol{\Lambda},\gamma)$ as diagonal matrix with entries $\mu(\lambda_i,\gamma)$ as mean of the spike eigenvalue distribution, F_γ as the Marčenko-Pastur distribution and s_c the value in $\mathbb{R}$ which satisfies $\max\left(\frac{k-d}{p-d}, 0\right) = \int_{s_c}^{(1+\sqrt{\gamma})^2} dF_\gamma(s)$.

Theorem 2 (Asymptotic PCR risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0,\infty)$, the expected risk of PCR will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\theta}})\right] \to \text{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 + \text{Var}_\gamma(\hat{\boldsymbol{\theta}}) + \sigma_\nu^2.$$

Proof. In the following, we split the proof into the bias-squared term and the variance term.

Bias-squared term. Let us start with the bias-squared term from (4) given by $\text{Bias}(\hat{\boldsymbol{\theta}})^2 = \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta}$, and with $\boldsymbol{\beta} = r_w \boldsymbol{W}\boldsymbol{C}_z(\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_z)^{-1}\boldsymbol{\theta}$ from the equivalence of the spiked covariance model to a data generator directly between features and outcomes. Further, we have that $\boldsymbol{W} = \boldsymbol{V}_d$ with the eigenvectors defined as $\boldsymbol{V} = \boldsymbol{I}_p$. Hence, since $\boldsymbol{V}_d \in \mathbb{R}^{p\times d}$, we know that the last $p - d$ rows will be zeros only. Therefore, we can write $\boldsymbol{W} = \boldsymbol{V}_d = \begin{bmatrix}\boldsymbol{V}_{d,d}\\ 0\end{bmatrix}$ where $\boldsymbol{V}_{d,d} \in \mathbb{R}^{d\times d}$ are the eigenvectors truncated to the first d-elements. Further, we use $\bar{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta}$ to write

$$\text{Bias}(\hat{\boldsymbol{\theta}})^2 = \bar{\boldsymbol{\beta}}^\top \begin{bmatrix}\boldsymbol{V}_{d,d}^\top & 0\end{bmatrix} \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi} \begin{bmatrix}\boldsymbol{V}_{d,d}\\ 0\end{bmatrix} \bar{\boldsymbol{\beta}} \tag{C-41}$$

Using $\Pi = (\boldsymbol{I}_p - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top)$ and $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ to expand $\boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}$ we obtain

$$\begin{aligned}\text{Bias}(\hat{\boldsymbol{\theta}})^2 &= \bar{\boldsymbol{\beta}}^\top \begin{bmatrix}\boldsymbol{V}_{d,d}^\top & 0\end{bmatrix} \left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top - \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top + \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\right) \begin{bmatrix}\boldsymbol{V}_{d,d}\\ 0\end{bmatrix} \bar{\boldsymbol{\beta}} && \text{(C-42)}\\ &= \bar{\boldsymbol{\beta}}^\top \left(\boldsymbol{\Lambda}_d - \boldsymbol{\Lambda}_d \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} - \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d}\boldsymbol{\Lambda}_d + \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}^\top \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d}\right) \bar{\boldsymbol{\beta}} && \text{(C-43)}\end{aligned}$$

where $\boldsymbol{\Lambda}_d$ are the first d eigenvalues and $\hat{\boldsymbol{V}}_d \in \mathbb{R}^{d\times k}$ are the first k estimated eigenvectors truncated to the first d elements in each vector. The latter happens due to the multiplication with the $\begin{bmatrix}\boldsymbol{V}_{d,d}^\top & 0\end{bmatrix}$ from left and its transpose from right. For the last term, we can expand $\boldsymbol{\Lambda} = \boldsymbol{I}_p + r_w^2 \boldsymbol{W}\boldsymbol{C}_z\boldsymbol{W}^\top$. Then we obtain

$$\begin{aligned}\boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}^\top \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} &= \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}^\top \boldsymbol{V}(\boldsymbol{I}_p + r_w^2 \boldsymbol{W}\boldsymbol{C}_z\boldsymbol{W}^\top)\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} &\text{(C-44)}\\ &= \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} + \boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d}(r_w^2 \boldsymbol{C}_z)\boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d} &\text{(C-45)}\end{aligned}$$

With this expression we observe that all terms in (C-43) contain the term $\boldsymbol{V}_{d,d}^\top \hat{\boldsymbol{V}}_d \hat{\boldsymbol{V}}_d^\top \boldsymbol{V}_{d,d}$ which is the same expression as we have for the generalization of the eigenvector shift in (5). Hence, for $p, n \to \infty$ such that $p/n \to \gamma$, we obtain

$$\mathrm{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 \to \bar{\boldsymbol{\beta}}^\top \left(\boldsymbol{\Lambda}_d - \boldsymbol{\Lambda}_d \boldsymbol{P}_k - \boldsymbol{P}_k\boldsymbol{\Lambda}_d + \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_z \boldsymbol{P}_k\right)\bar{\boldsymbol{\beta}} \tag{C-46}$$

which is the expression for the bias-squared term.

Variance term. Let us start with the variance term from (4) given by $\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n}\operatorname{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}}\right)$. Here, we follow a similar approach by expanding the covariance terms and exploiting the multiplication of eigenvectors. For the trace term, we thus obtain

$$\begin{aligned}\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}} &= \hat{\boldsymbol{V}}^\top \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}^{-1} &\text{(C-47)}\\ &= \hat{\boldsymbol{V}}^\top \boldsymbol{V}\left(\boldsymbol{W} r_w^2 \boldsymbol{C}_z \boldsymbol{W}^\top + \boldsymbol{I}_p\right)\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}^{-1} &\text{(C-48)}\\ &= \hat{\boldsymbol{V}}^\top \boldsymbol{V}\left(\begin{bmatrix}\boldsymbol{V}\\ 0\end{bmatrix} r_w^2 \boldsymbol{C}_z \begin{bmatrix}\boldsymbol{V}^\top & 0\end{bmatrix} + \boldsymbol{I}_p\right)\boldsymbol{V}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}^{-1} &\text{(C-49)}\\ &= \left(\hat{\boldsymbol{V}}^\top \boldsymbol{V}_d r_w^2 \boldsymbol{C}_z \boldsymbol{V}_d^\top \hat{\boldsymbol{V}} + \boldsymbol{I}_k\right)\hat{\boldsymbol{\Lambda}}^{-1} &\text{(C-50)}\end{aligned}$$

Here, we can notice that the expression $\hat{\boldsymbol{V}}^\top \boldsymbol{V}_d$ can give us part of the eigenvector shift equation from (5) to obtain

$$\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}} \to \left(\begin{bmatrix}\boldsymbol{P}_k^{1/2}\\ 0\end{bmatrix} r_w^2 \boldsymbol{C}_z \begin{bmatrix}\boldsymbol{P}_k^{1/2} & 0\end{bmatrix} + \boldsymbol{I}_k\right)\hat{\boldsymbol{\Lambda}}^{-1} \tag{C-51}$$

since both $\boldsymbol{C}_z$ and $\boldsymbol{P}_k$ are diagonal, we can write

$$\hat{\boldsymbol{V}}^\top \boldsymbol{C}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}} \to \begin{bmatrix}\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_d & 0\\ 0 & \boldsymbol{I}_{k-d}\end{bmatrix}\hat{\boldsymbol{\Lambda}}^{-1}. \tag{C-52}$$

Including this into the full variance term, and considering each term in the trace individually, we obtain

$$\mathrm{Var}_\gamma(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n}\left(\sum_{i=1}^{\min(d,k)}(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k)_{[i]}\frac{1}{\hat{\lambda}_i} + \sum_{i=\min(d,k)+1}^{k}\frac{1}{\hat{\lambda}_i}\right) \tag{C-53}$$

where $\boldsymbol{A}_{[i]}$ denotes the ith diagonal element of $\boldsymbol{A}$. Here, we notice that the first term corresponds to the spike eigenvalues since the sum goes only over the top $\min(d,k)$ eigenvalues and the second summation goes over the remaining eigenvalues including all noise terms. Hence, the first term can be written as

$$\sum_{i=1}^{\min(d,k)}(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k)_{[i]}\frac{1}{\hat{\lambda}_i} = \operatorname{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k)\frac{1}{\mu(\boldsymbol{\Lambda},\gamma)}\right] \tag{C-54}$$

where $\mu(\boldsymbol{\Lambda},\gamma)$ is a diagonal matrix with entries $\mu(\lambda_i,\gamma)$ as mean of the spike eigenvalue distribution. For the second term, we can write the sum over the eigenvalues $\hat{\lambda}$ as the integral over the spectral measure $F_{\hat{\boldsymbol{C}}_{MP}}$ of the covariance for the Marčenko-Pastur distribution $\hat{\boldsymbol{C}}_{MP}$.

$$\sum_{i=\min(d,k)+1}^{k}\frac{1}{\hat{\lambda}_i} = (p-d)\int_{s_c}^{(1+\sqrt{\gamma})^2}\frac{1}{s}dF_{\hat{\boldsymbol{C}}_{MP}}(s) \tag{C-55}$$

Since this term is only over the noise terms, the integral upper bound is given by the upper bound of the Marčenko-Pastur distribution $(1+\sqrt{\gamma})^2$. The integral lower bound s_c corresponds to the $p-k$ largest eigenvalue. There is a scaling factor of $p-d$ as this is the number of eigenvalues corresponding to the part of the eigenvalue distribution. Now, we know that in the limit $p, n \to \infty$ such that $p/n \to \gamma$ the spectral measure will almost surely converge to the Marčenko-Pastur distribution F_γ. Therefore we obtain

$$\sum_{i=\min(d,k)+1}^{k} \frac{1}{\hat{\lambda}_i} \to (p-d) \int_{s_c}^{(1+\sqrt{\gamma})^2} \frac{1}{s} dF_\gamma(s) \tag{C-56}$$

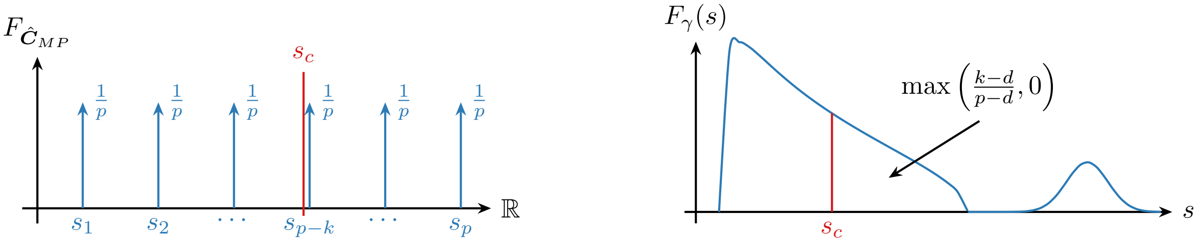
There are two steps to solve this integral. First, we need to find out the lower integral bound s_c and second, solve the integral itself. For $s_c = -\infty$, one can use the closed-form solution of the Stieltjes transformation $\varphi(z)$ of the Marčenko-Pastur distribution and evaluate it at $z = 0$. However, there is no known closed-form solution for general s_c. We therefore solve this part numerically.

Step 1 obtain the lower bound s_c: We can view the spectral measure as $F_{\hat{\boldsymbol{C}}_{MP}}$ as a series of $(p-d)$ impulses at s_i with magnitude $1/p$ (as the sum is normalized to 1). Since we only consider the $k-d$ largest eigenvalues (the remaining ones are considered in the first term of the trace for the spike part of the covariance), we know that their sum is $(k-d)/(p-d)$, see Figure C-4a. This sum is the same as the integral from s_c over the Marčenko-Pastur distribution, see Figure C-4b. Therefore we can find the lower integral bound s_c by solving the following numerically

$$\max\left(\frac{k-d}{p-d}, 0\right) = \int_{sc}^{(1+\sqrt{\gamma})^2} dF_\gamma(s). \tag{C-57}$$

Here, the $\max$ is necessary as we could have $k \leq d$. Then s_c will become $(1+\sqrt{\gamma})^2$ which means that the Marčenko-Pastur part will not contribute to the variance term.

Step 2 solve integral of interest: Given the lower bound s_c, we can solve the integral numerically from s_c to the upper bound $(1+\sqrt{\gamma})^2$. Therefore, we obtain a solution for the second term, which is not based on data but on the properties of our data matrix, especially γ, d and k. This concludes the full proof for the asymptotics of the parameter norm. □



(a)

(b)

Figure C-4: *(a)* spectral measure impulses and lower integral bound of integral s_c. *(b)* Illustration of spiked covariance distribution for $\gamma = 0.3$ with specific lower integration bound.

C.2.2. Risk of baseline methods

Let us first re-state the risk for full regression. Note that this is a known result from Hastie et al. (2022, Lemma 1). Hence, we refer to the original reference for the proof.

$$\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\beta}})\right] = \boldsymbol{\beta}^\top \boldsymbol{\Pi} \boldsymbol{C} \boldsymbol{\Pi} \boldsymbol{\beta} + \frac{\sigma_\nu^2}{n} \operatorname{Tr}(\boldsymbol{C}\hat{\boldsymbol{C}}^{-1}) + \sigma_\nu^2. \tag{8}$$

Asymptotic full regression risk. For the asymptotic full regression risk, we have $k = p$. We first re-state the theorem:

Theorem 3 (Asymptotic full regression risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0, \infty)$, the expected risk of the full regression model will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\beta}})\right] \to \mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2 + \mathrm{Var}_\gamma(\hat{\boldsymbol{\beta}}) + \sigma_\nu^2$$

with the asymptotic squared bias term as $\mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2 = 0$ for $\gamma < 1$ and $\mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2$ as in Theorem 2 with $k = p$ for $\gamma \geq 1$; and the variance term $\mathrm{Var}_\gamma(\hat{\boldsymbol{\beta}})$ equal to the definition of the variance in Theorem 2 with $k = p$.

Proof. For the *asymptotic variance* term, we cannot simplify the results from Theorem 2 except for the case that $\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top = \boldsymbol{I}$ here because the eigenvectors are not truncated.

For the *asymptotic bias-squared* term, we know in the full regression model that $\boldsymbol{\Pi}$ is a projection matrix onto the null space of $\boldsymbol{X}$. Hence, we have equally to the asymptotic result from Hastie et al. (2022, Theorem 1) that

$$\mathrm{Bias}_\gamma(\hat{\boldsymbol{\beta}})^2 \to \begin{cases} 0 & \gamma < 1 \\ \boldsymbol{\beta}^\top \boldsymbol{\Pi}\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\beta} & \gamma \geq 1 \end{cases} \tag{C-58}$$

which concludes the proof as we cannot simplify the expression further, given the non-isotropic covariance matrix. □

C.3. Analysis under covariate shift

C.3.1. Risk of PCR

Expected covariate shifted risk of PCR. Let us first re-state the expected risk result for PCR under covariate shift:

$$\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] = (\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta}) + \frac{\sigma_\nu^2}{n}\,\mathrm{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}_S^{-1}\hat{\boldsymbol{V}}\right) + \sigma_T^2, \tag{9}$$

with $\boldsymbol{\beta}_T = r_w \boldsymbol{W}\boldsymbol{C}_{z,T}(\boldsymbol{I}_d + r_w^2 \boldsymbol{C}_{z,T})^{-1}\boldsymbol{\theta}$ and $\sigma_T^2 = \sigma_\varepsilon^2 + \boldsymbol{\theta}^\top(\boldsymbol{I}_d + r_w^2\boldsymbol{C}_{z,T})^{-1}\boldsymbol{C}_{z,T}\boldsymbol{\theta}$. (9)

Proof. First, we estimate the parameters which is done using source/training data which is equal to the non-covariate shifted case. Then, we define the risk as the expectation over the mean squared error over test data and follow a similar derivation as for (4):

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{(\boldsymbol{x}_T, y_T)\sim\mathcal{D}_T}\left[(y_T - \hat{y}_T(\boldsymbol{x}_T))^2\right] && \text{(C-59)} \\ &= \mathbb{E}_{\boldsymbol{x}_T}\left[(\boldsymbol{\beta}_T^\top \boldsymbol{x}_T + \nu_T - \hat{y}_T(\boldsymbol{x}_T))^2\right] && \text{(C-60)} \\ &= \mathbb{E}_{\boldsymbol{x}_T}\left[\left((\boldsymbol{\beta}_T - \hat{\boldsymbol{V}}\hat{\boldsymbol{\theta}})^\top \boldsymbol{x}_T + \nu_T\right)^2\right] && \text{(C-61)} \\ &= (\boldsymbol{\beta}_T - \hat{\boldsymbol{V}}\hat{\boldsymbol{\theta}})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \hat{\boldsymbol{V}}\hat{\boldsymbol{\theta}}) + \nu_T\nu_T^\top && \text{(C-62)} \end{aligned}$$

We consider again the orthogonal projector $\boldsymbol{\Phi} = \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top$ to rewrite the following:

$$\boldsymbol{\beta}_t - \hat{\boldsymbol{V}}\hat{\boldsymbol{\theta}} = \boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta} - \hat{\boldsymbol{V}}\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{U}}^\top \nu. \tag{C-63}$$

Finally, using the previous two results following the same derivation as for the non-covariate shifted risk, we can write the expected risk w.r.t. the noise as

$$\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] = (\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta}) + \frac{\sigma_\nu^2}{n}\,\mathrm{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \hat{\boldsymbol{C}}_S^{-1}\hat{\boldsymbol{V}}\right) + \sigma_T^2, \tag{C-64}$$

which concludes the proof. □

Covariate shifted asymptotic PCR risk. In order to prove Theorem 4, we re-state the results for the bias-squared and the variance terms first: The *asymptotic bias-squared* term is given by

$$\mathrm{Bias}_{\gamma,T}(\hat{\boldsymbol{\theta}})^2 = [\bar{\boldsymbol{\beta}}_T^\top \bar{\boldsymbol{\beta}}^\top] \begin{bmatrix} \boldsymbol{\Lambda}_{d,T} & -\boldsymbol{\Lambda}_{d,T}\boldsymbol{P}_k \\ -\boldsymbol{P}_k\boldsymbol{\Lambda}_{d,T} & \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T}\boldsymbol{P}_k \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\beta}}_T \\ \bar{\boldsymbol{\beta}} \end{bmatrix} \tag{10}$$

with $\bar{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}}_T = \boldsymbol{W}^{-1}\boldsymbol{\beta}_T$ and $\boldsymbol{\Lambda}_{d,T} \in \mathbb{R}^{d\times d}$ as the truncation of $\boldsymbol{\Lambda}_T$ to the first d dimensions. The *asymptotic variance* term is

$$\mathrm{Var}_{\gamma,T}(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n}\left(\mathrm{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T} + \boldsymbol{I}_k)\frac{1}{\mu(\boldsymbol{\Lambda},\gamma)}\right] + (p-d)\int_{s_c}^{(1+\sqrt{\gamma})^2}\frac{1}{s}dF_\gamma(s)\right) \tag{11}$$

with $\mu(\boldsymbol{\Lambda},\gamma)$ as diagonal matrix with entries $\mu(\lambda_i,\gamma)$ as mean of the spike eigenvalue distribution, F_γ as the Marčenko-Pastur distribution and s_c the value in $\mathbb{R}$ which satisfies $\max\left(\frac{k-d}{p-d},0\right) = \int_{s_c}^{(1+\sqrt{\gamma})^2} dF_\gamma(s)$.

Theorem 4 (Covariate-shifted asymptotic PCR risk)**.** In the asymptotic limit $n,p\to\infty$, such that $\frac{p}{n}\to\gamma\in(0,\infty)$, the expected risk of PCR under covariate shift will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}_T}\left[R(\hat{\boldsymbol{\theta}})\right] \to \mathrm{Bias}_{\gamma,T}(\hat{\boldsymbol{\theta}})^2 + \mathrm{Var}_{\gamma,T}(\hat{\boldsymbol{\theta}}) + \sigma_T^2.$$

Proof. We split the proof into the bias-squared and variance term.

Bias-squared term. We start with the bias-squared term $\mathrm{Bias}_T(\boldsymbol{\beta})^2 = (\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})$ from (9) and multiply out the terms while using the definition of $\boldsymbol{\beta} = \boldsymbol{W}\bar{\boldsymbol{\beta}}$ and similarly for the $\boldsymbol{\beta}_T$

$$\begin{aligned}\mathrm{Bias}_T(\boldsymbol{\beta})^2 =& \bar{\boldsymbol{\beta}}_T^\top \boldsymbol{W}^\top \boldsymbol{C}_T \boldsymbol{W}\bar{\boldsymbol{\beta}}_T\\ &-\bar{\boldsymbol{\beta}}_T^\top \boldsymbol{W}^\top \boldsymbol{C}_T \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \boldsymbol{W}\bar{\boldsymbol{\beta}}\\ &-\bar{\boldsymbol{\beta}}^\top \boldsymbol{W}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T \boldsymbol{W}\bar{\boldsymbol{\beta}}_T\\ &+\bar{\boldsymbol{\beta}}^\top \boldsymbol{W}^\top \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top \boldsymbol{W}\bar{\boldsymbol{\beta}}\end{aligned} \tag{C-65}$$

where we can use the same results about the asymptotic eigenvector shifts $\boldsymbol{P}_k$ when expanding the covariance $\boldsymbol{C}_T = \boldsymbol{U}\boldsymbol{\Lambda}_T\boldsymbol{V}^T$. Then, we can summarize the terms into

$$\mathrm{Bias}_{\gamma,T}(\boldsymbol{\beta})^2 \to \begin{bmatrix}\bar{\boldsymbol{\beta}}_T^\top & \bar{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{d,T} & -\boldsymbol{\Lambda}_{d,T}\boldsymbol{P}_k\\ -\boldsymbol{P}_k\boldsymbol{\Lambda}_{d,T} & \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T}\boldsymbol{P}_k\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{\beta}}_T\\ \bar{\boldsymbol{\beta}}\end{bmatrix} \tag{C-66}$$

Which yields the result for the bias-squared term.

Variance term. We start with the variance term $\mathrm{Var}_T(\boldsymbol{\beta}) = \frac{\sigma_\nu^2}{n}\mathrm{Tr}\left(\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\hat{\boldsymbol{C}}_S^{-1}\hat{\boldsymbol{V}}\right)$ from (9). Let us focus on the Trace part first

$$\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}^\top\boldsymbol{V}\boldsymbol{\Lambda}_T\boldsymbol{V}^\top\hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}^{-1} \tag{C-67}$$

$$= \hat{\boldsymbol{V}}^\top\boldsymbol{V}\left(\boldsymbol{W}r_w^2\boldsymbol{C}_{z,T}\boldsymbol{W}^\top + \boldsymbol{I}_p\right)\boldsymbol{V}^\top\hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}^{-1} \tag{C-68}$$

$$= \left(\hat{\boldsymbol{V}}^\top\boldsymbol{V}_d r_w^2\boldsymbol{C}_{z,T}\boldsymbol{V}_d^\top\hat{\boldsymbol{V}} + \boldsymbol{I}_k\right)\hat{\boldsymbol{\Lambda}}^{-1} \tag{C-69}$$

Again, we utilize the results from the eigenvector shift equation from (5) to obtain

$$\hat{\boldsymbol{V}}^\top \boldsymbol{C}_T\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\hat{\boldsymbol{C}}^{-1}\hat{\boldsymbol{V}} \to \left(\begin{bmatrix}\boldsymbol{P}_k^{1/2}\\ 0\end{bmatrix} r_w^2\boldsymbol{C}_{z,T}\begin{bmatrix}\boldsymbol{P}_k^{1/2} & 0\end{bmatrix} + \boldsymbol{I}_k\right)\hat{\boldsymbol{\Lambda}}^{-1} \tag{C-70}$$

Here, we notice that the main difference to the non-covariate shifted result lies in the first part where we use $\boldsymbol{C}_{z,T}$ instead of $\boldsymbol{C}_z$. Therefore, using the same arguments as in the proof for Theorem 2 we obtain the final result for the variance term as

$$\mathrm{Var}_{\gamma,T}(\boldsymbol{\beta}) \to \frac{\sigma_\nu^2}{n}\left(\mathrm{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_{z,T} + \boldsymbol{I}_k)\frac{1}{\mu(\boldsymbol{\Lambda},\gamma)}\right] + (p-d)\int_{s_c}^{(1+\sqrt{\gamma})^2}\frac{1}{s}dF_\gamma(s)\right) \tag{C-71}$$

which concludes the proof. □

C.3.2. Risk of baseline methods

Theorem 5 (Covariate-shifted asymptotic full regression risk)**.** In the asymptotic limit $n, p \to \infty$, such that $\frac{p}{n} \to \gamma \in (0, \infty)$, the expected risk of the full regression model under covariate shift will converge almost surely to

$$\mathbb{E}_{\boldsymbol{\nu}}\left[R(\hat{\boldsymbol{\beta}})\right] \to \mathrm{Bias}_{\gamma,T}(\hat{\boldsymbol{\beta}})^2 + \mathrm{Var}_{\gamma,T}(\hat{\boldsymbol{\beta}}) + \sigma_T^2$$

with the asymptotic squared bias term as

$$\mathrm{Bias}_{\gamma,T}(\hat{\boldsymbol{\beta}})^2 = (\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}})^\top \boldsymbol{\Lambda}_{d,T}(\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}})$$

for $\gamma < 1$ and as in Theorem 4 for $\gamma \geq 1$ with $k = p$. The variance term $Var_\gamma(\hat{\boldsymbol{\beta}})$ is equal to the definition of the variance in Theorem 4 with $k = p$.

Proof. The variance term and the bias-squared term for $\gamma \geq 1$ are equal to the proof for Theorem 4. For the bias-square term for $\gamma < 1$ we have that $\mathrm{Bias}_T(\boldsymbol{\beta})^2 = (\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{C}_T(\boldsymbol{\beta}_T - \boldsymbol{\Phi}\boldsymbol{\beta})$. Now, we can follow our derivation for the bias-squared term of Theorem 2 with $\boldsymbol{P}_k = \boldsymbol{I}$ since we do not truncate the eigenvectors $\hat{\boldsymbol{V}}$ in this full regression case. Hence, we expand $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ to obtain

$$\mathrm{Bias}_{\gamma,T}(\hat{\boldsymbol{\beta}})^2 \to (\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}})^\top \boldsymbol{\Lambda}_{d,T}(\bar{\boldsymbol{\beta}}_T - \bar{\boldsymbol{\beta}}) \tag{C-72}$$

since $\boldsymbol{\Phi} = \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top = \boldsymbol{I}$ for $\gamma < 1$. This concludes the proof. □