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

UPPSALA UNIVERSITET

Daniel Gedon, Antônio H. Ribeiro, Thomas B. Schön

Uppsala University (Sweden) *contact: daniel.gedon@it.uu.se*

Background

Spiked covariance model. Population covariance $\boldsymbol{C}$ of $\boldsymbol{x}$, eigenvectors $\boldsymbol{v}$, eigenvalues λ,

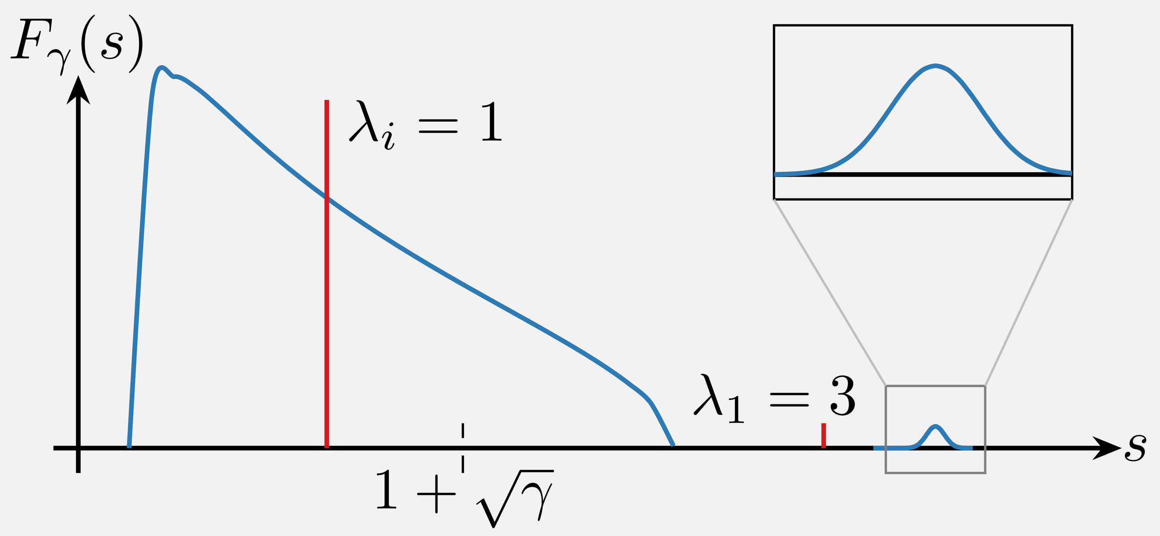
$$\boldsymbol{C} = \boldsymbol{I} + \sum_{i=1}^{d} \boldsymbol{v}_i \lambda_i \boldsymbol{v}_i^\top .$$

Eigenvalue shift. Distribution of spike sample eigenvalue $\hat{\lambda}_1$, depending on λ_1:

- $\lambda_1 \in [1, 1+\sqrt{\gamma}]$: spike at $\mu(\gamma) = (1+\sqrt{\gamma})^2$.
- $\lambda_1 > 1+\sqrt{\gamma}$: spike Normal with $\mu(\lambda, \gamma) = \gamma \frac{\lambda}{\lambda-1} + \lambda$.

Eigenvector shift. Sample eigenvectors $\hat{\boldsymbol{v}}$. As $p/n \to \gamma$,

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



[1] Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. Annals of Statistics, 2001.
[2] Johnstone, I. M. and Paul, D. PCA in high dimensions: An orientation. Proceedings of the IEEE, 2018.

Theory – preliminary

Risk. $R(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{(\boldsymbol{x}_0, y_0)\sim\mathcal{D}} \left[(y_0 - \hat{y}_0(\boldsymbol{x}_0))^2 \right]$

Eigenvector shift. Generalisation to all eigenvectors

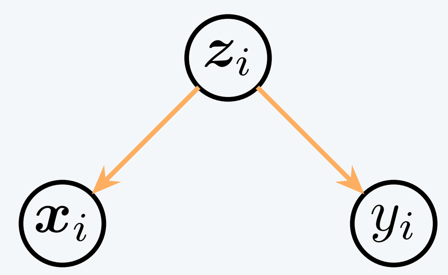
$$\boldsymbol{P}_k = \begin{cases} \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, ..., (\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k)^2, 0, ..., 0\right), & k < d, \\ \operatorname{diag}\left((\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1)^2, ..., (\boldsymbol{v}_d^\top \hat{\boldsymbol{v}}_d)^2\right),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

Motivation

- Overparameterized models, $\gamma = \frac{p}{n} > 1 \to$ double descent
- Real-world data often on a low-dimensional manifold
- PCR (= PCA + linear regression) widely adopted in practice

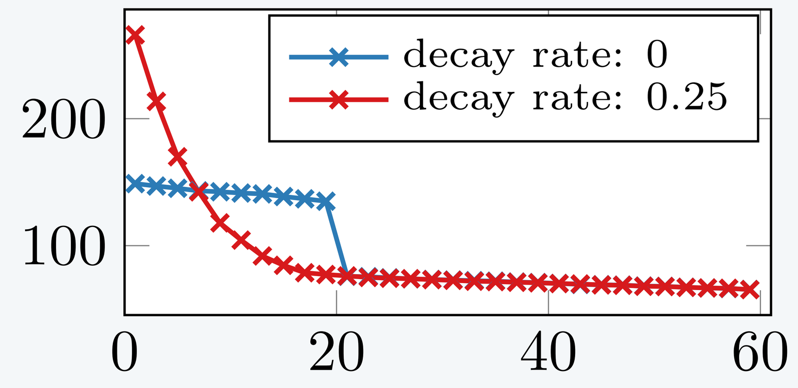
[3] Belkin, M., Hsu, D. J., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. PNAS, 2019

Data generator



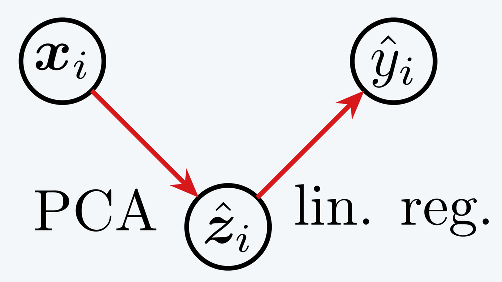
$$\boldsymbol{x}_i = \boldsymbol{W} r_w \boldsymbol{z}_i + \boldsymbol{e}_i,$$
$$y_i = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \varepsilon_i.$$

Eigenvalues of $\boldsymbol{x}$.



Dimensions: $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{z} \in \mathbb{R}^d$, $\hat{\boldsymbol{z}} \in \mathbb{R}^k$.

Model



SVD: $\boldsymbol{X} \approx \hat{\boldsymbol{U}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{V}}^\top$,

PCA: $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{V}}^\top \boldsymbol{x}_i$,

lin. reg. $\hat{y}_i = \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{z}}_i$.

Theory

Theorem. Let $n, p \to \infty$ with $\frac{p}{n} \to \gamma \in (0, \infty)$, we have a.s.

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

$$\operatorname{Bias}_\gamma(\hat{\boldsymbol{\theta}})^2 = \bar{\boldsymbol{\beta}}^\top \big(\boldsymbol{\Lambda}_d - \boldsymbol{\Lambda}_d \boldsymbol{P}_k - \boldsymbol{P}_k \boldsymbol{\Lambda}_d + \boldsymbol{P}_k + \boldsymbol{P}_k r_w^2 \boldsymbol{C}_z \boldsymbol{P}_k \big) \bar{\boldsymbol{\beta}},$$

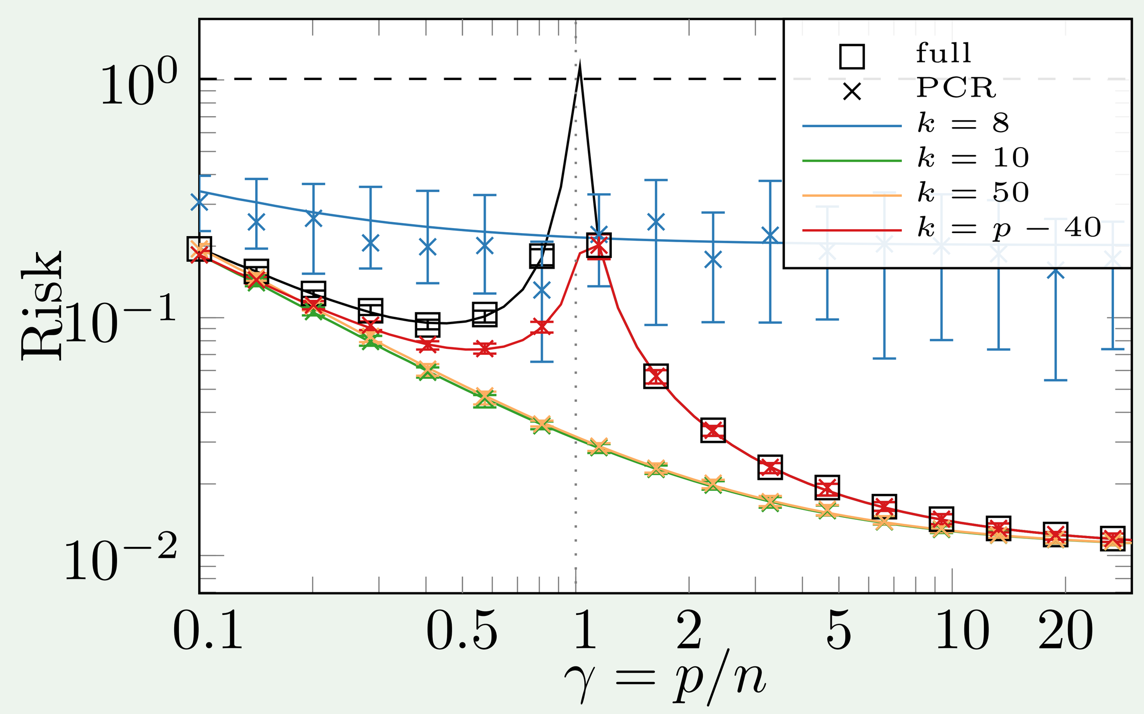
$$\operatorname{Var}_\gamma(\hat{\boldsymbol{\theta}}) = \frac{\sigma_\nu^2}{n} \left(\operatorname{Tr}\left[(\boldsymbol{P}_k r_w^2 \boldsymbol{C}_z + \boldsymbol{I}_k) \frac{1}{\mu(\boldsymbol{\Lambda}, \gamma)} \right] + (p-d) \int_{s_c}^{(1+\sqrt{\gamma})^2} \frac{1}{s} dF_\gamma(s) \right).$$

Interpretation.

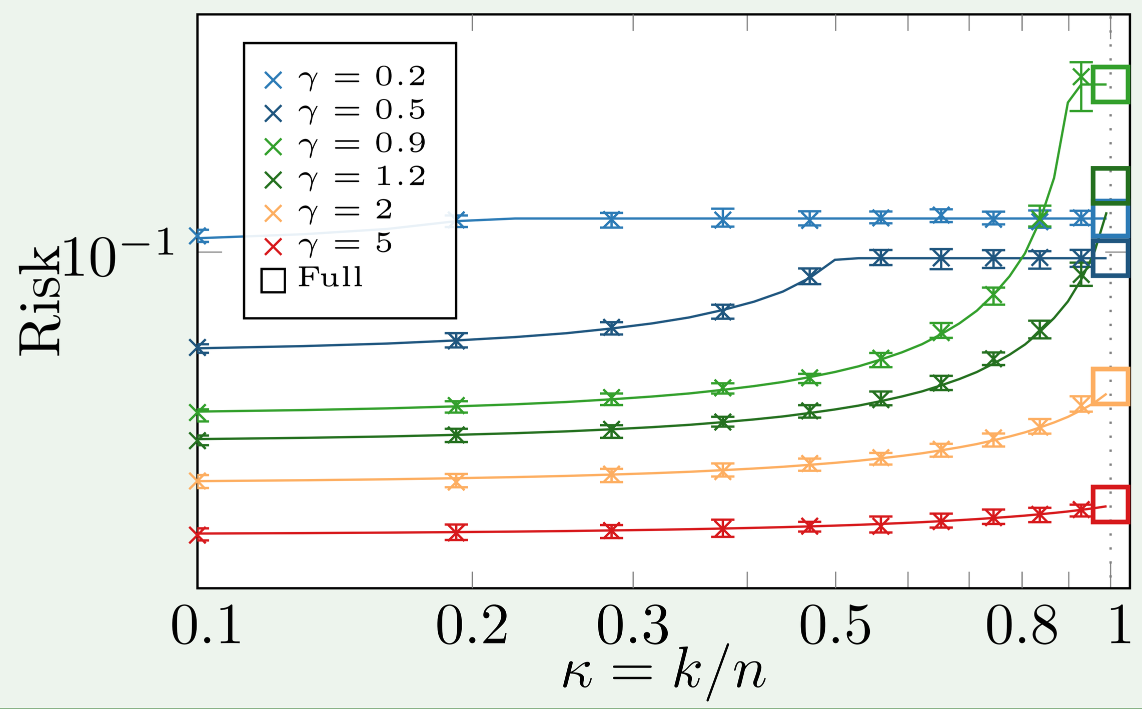
1. *Bias* as scaled d-dimensional subspace of eigenvalues $\boldsymbol{\Lambda}$.
2. *Variance* with $k \le d \to s_c = (1+\sqrt{\gamma})^2$ $\to$ integral term= 0.
3. Principal components $k \to$ cut-off for considered data distribution.

Results

Main result. Number of spikes $d = 10$.



PCA-projection space.



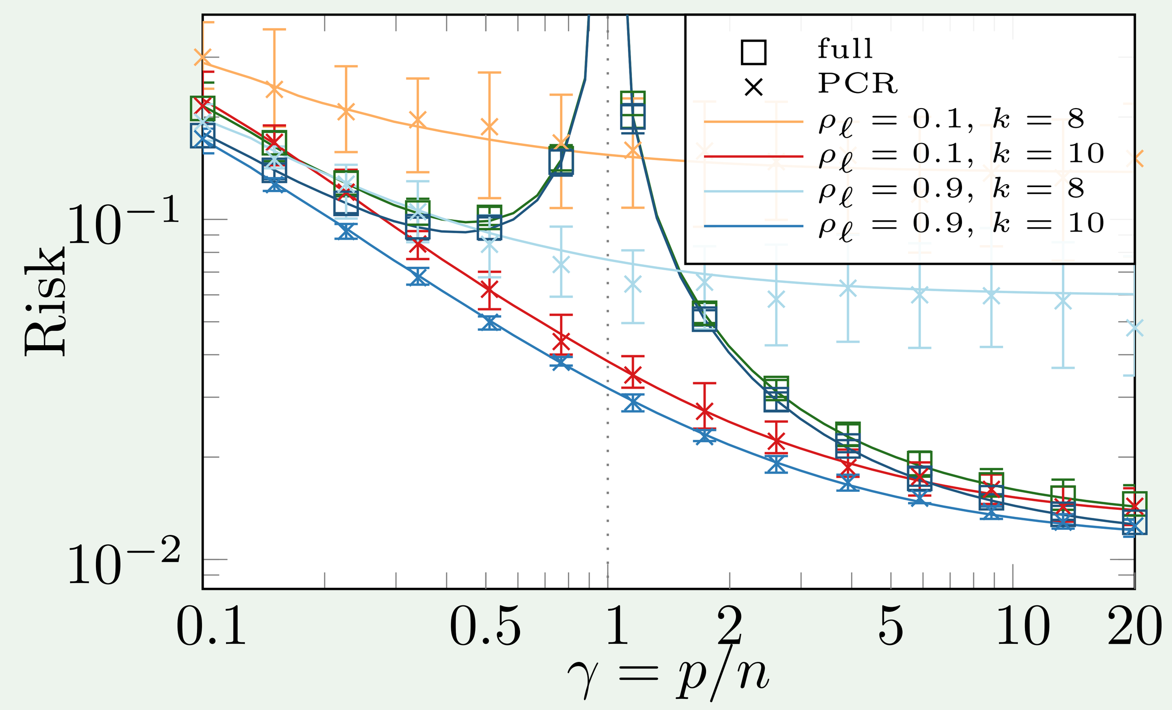
Covariate-shift

Assumption. Train/source $\boldsymbol{C}_S = \boldsymbol{V} \boldsymbol{\Lambda}_S \boldsymbol{V}^\top$ and test covariance $\boldsymbol{C}_T = \boldsymbol{V} \boldsymbol{\Lambda}_T \boldsymbol{V}^\top$.

Theorem. Similar to in-distribution risk but scaled. But combinations of $\boldsymbol{C}_S$, $\boldsymbol{C}_T$ here.

Results.

Data generator with $\operatorname{Cov}(\boldsymbol{v}_S, \boldsymbol{v}_T) = \sigma_\ell^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.



Observation.

1. Same behaviour as for in-distribution data.
2. Higher correlation $\rho \to$ lower risk.

Conclusion

Summary.

- Asymptotic risk of PCR under spiked covariance model. Tool: random matrix theory.
- Guarantee: widely used model & real-world data structures

Guide to choosing k.

- High k increases variance contribution.
- For $0.5 < \gamma < 2$ a suitable k has little effect.

Limitations.

- Linear supervised setting.
- Asymptotic results — no finite sample guarantees.