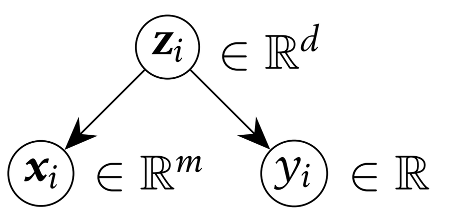
No Double Descent in PCA: Training and Pre-Training in High Dimensions

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

Department of Information Technology, Uppsala University, Sweden

Problem formulation

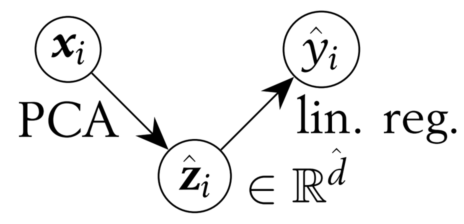
Data generator



$x_i = Dz_i + e_i$
$y_i = \theta z_i + \varepsilon_i$
$i = 1, \dots, n$

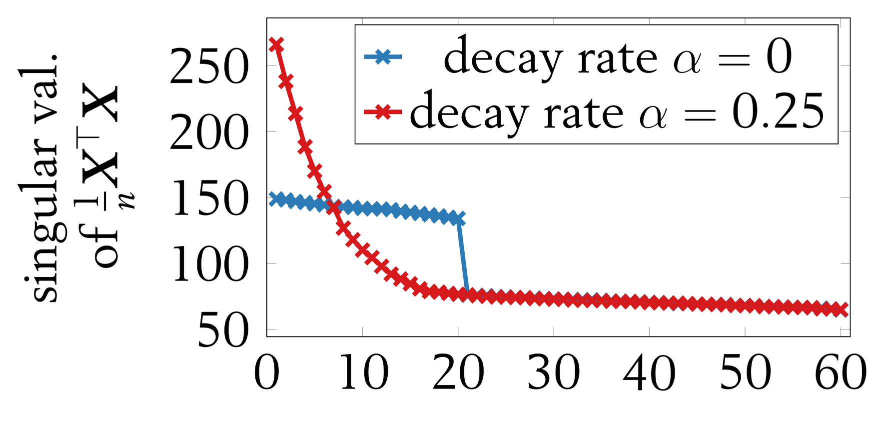
$z_i \sim \mathcal{N}(0, \lambda_i^2 I_d)$ with $\lambda_i^2 = \exp(-i\alpha)$

Model PCA-regression



PCA: $\hat{z}_i = \hat{V}^\top x_i, \quad X \approx \hat{U}\hat{\Sigma}\hat{V}^\top$

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



Motivation

- Realistic data on low-dim. manifold.
- PCA-regression similar in structure to successful encoder-decoder.

Aim: Understand the model generalization in high dimensions.

Supervised case – Analysis

Analyze risk on new data: $R(\hat{\theta}) = \mathbb{E}_{y_0}\left[(y_0 - \hat{y}_0)^2\right]$

Lemma Sample covariance $\hat{C} = \frac{1}{n}X^\top X$ and the true covariance C. Orthogonal projectors $\Pi = I_m - \hat{V}\hat{V}^\top$. Then,

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

Compare with Hastie et al. [1] for direct linear regression:

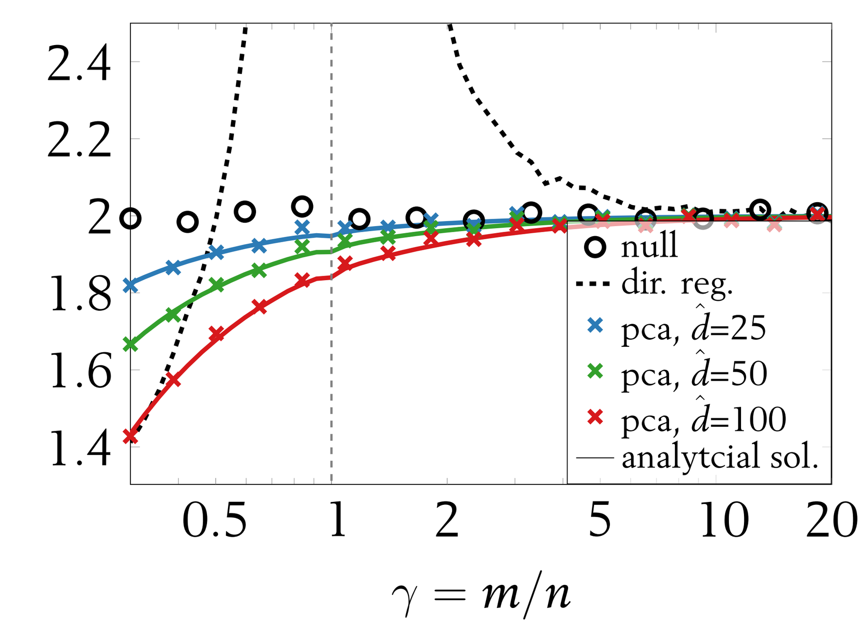
$$\mathbb{E}_\epsilon\left[R(\hat{\theta})\right] = \beta^\top \Pi C \Pi \beta + \frac{\sigma_\epsilon^2}{n}\operatorname{Tr}(C\hat{C}^+) + \sigma_\epsilon^2$$
$$= \text{bias}^2 + \text{variance} + \text{irreducible noise}.$$

Interpretation:

- Only variance term differs
- Estimated eigenvectors $\hat{V}$ project covariance C into $\hat{d}$-dimensional subspace → expect no interpolation peak at $\gamma = 1$

Supervised case – Numerical results

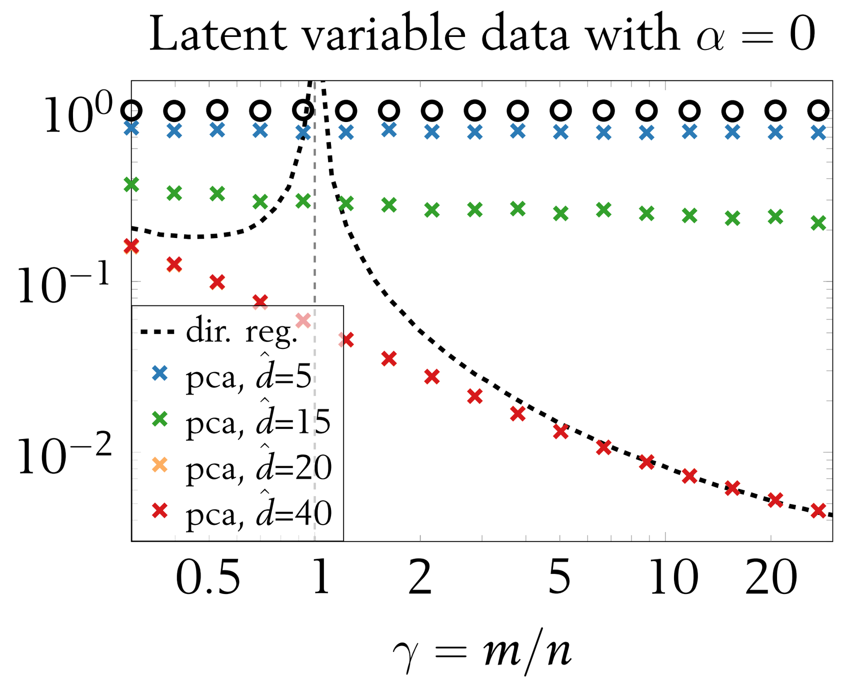
Isotropic data

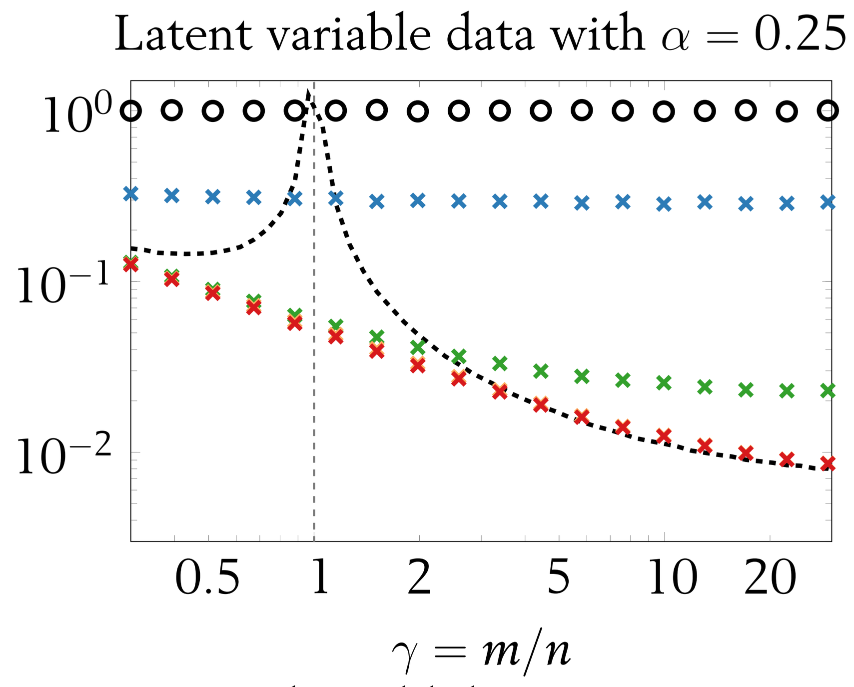


Isotropic setup:
Use $n = 400$ samples, $C = I_m \to d = m$.

Interpretation

1. Numerical simulation and analytical solution align.
2. No interpolation peak at $\gamma = 1$.

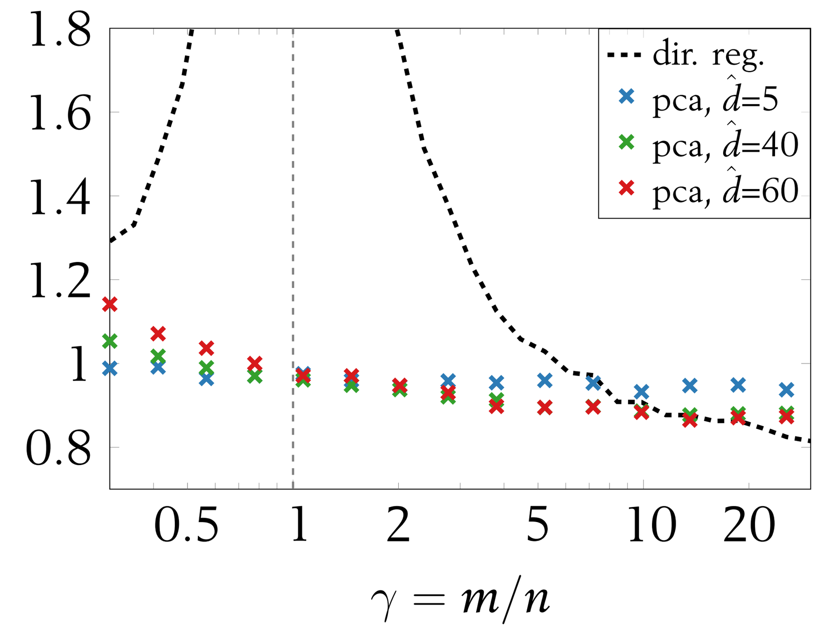




Latent var. setup: $d = 20$, $n = 400$.

Interpretation

1. $\hat{d} \geq d \to$ PCA-regression=dir. reg. for γ large/small.
2. $\hat{d} < d \to$ solution is suboptimal.



Genetics data

- Predict phenotypes from 1.1M genotypes.
- Resemblance to the latent variable results.

Pre-training the PCA – Setup

Two step training procedure:

1. Pre-training data set $\{x_i\}_{i=1}^{n_p}$ → unsupervised pre-training of PCA.
2. Training data set $\{x_i, y_i\}_{i=1}^{n}$ → linear regression on the PCA features $\hat{z}_i$.

⇒ Setting comparable to pre-training of encoder-decoder models.

For technical reasons: orthogonalize features and noise $x_i = Dz_i + D_\perp e_i$. Then:

Model: $\hat{z}_i = \hat{V}^\top x_i,$

Data generator: $z_i = D^+(x_i - D_\perp e_i) = D^+ x_i.$

Interpretation Correct estimation of true eigenvectors D^+ with $\hat{V}$ crucial.

Pre-training – Analysis

Define projection loss: $\mathcal{L}(D) = \mathbb{E}\left[\|x\|_2^2 - \|D^+x\|_2^2\right]; \quad \mathcal{L}(\hat{V}) = \mathbb{E}\left[\|x\|_2^2 - \|\hat{V}^\top x\|_2^2\right]$

Theorem Take $t > 0$, $k_j^2 = s_j(s_j + \operatorname{Tr}(C))$, then

$$P\left(\mathcal{L}(\hat{V}) - \mathcal{L}(D) > t\right) \leq$$
$$\leq \frac{4}{t\,n_p}\left(\sum_{i=1}^{\min(d,\hat{d})}\sum_{j=i+1}^{m}\frac{k_j^2}{|s_i - s_j|} + \sum_{i=\hat{d}}^{d}\sum_{j=1}^{m}\frac{k_j^2 s_i}{(s_i - s_j)^2} + \sum_{i=d}^{\hat{d}}\sum_{j=1}^{m}\frac{k_j^2 s_j}{(s_i - s_j)^2}\right).$$

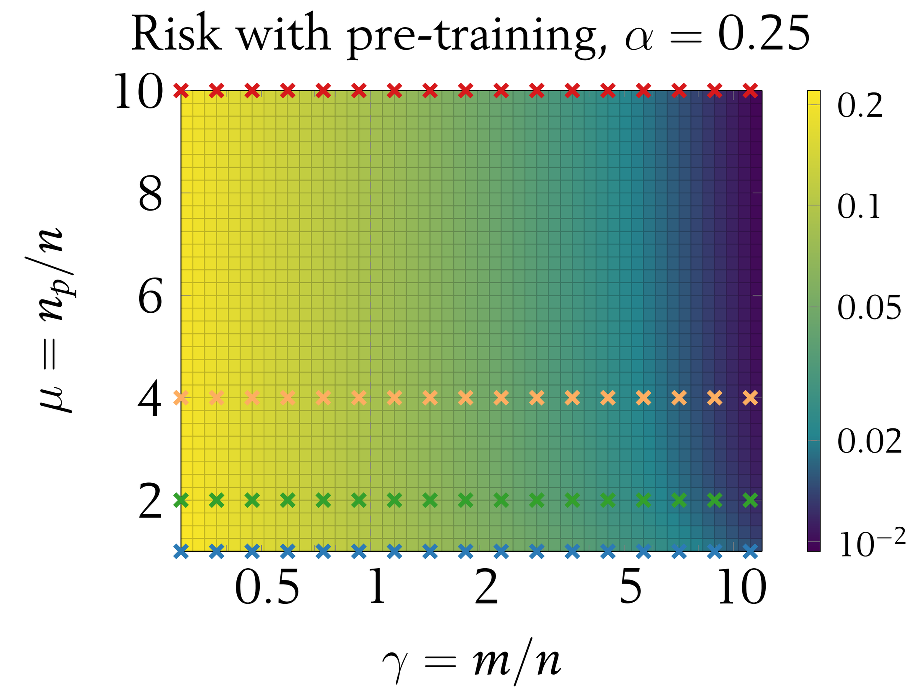
Interpretation Good covariance estimation $\hat{V}$ if:

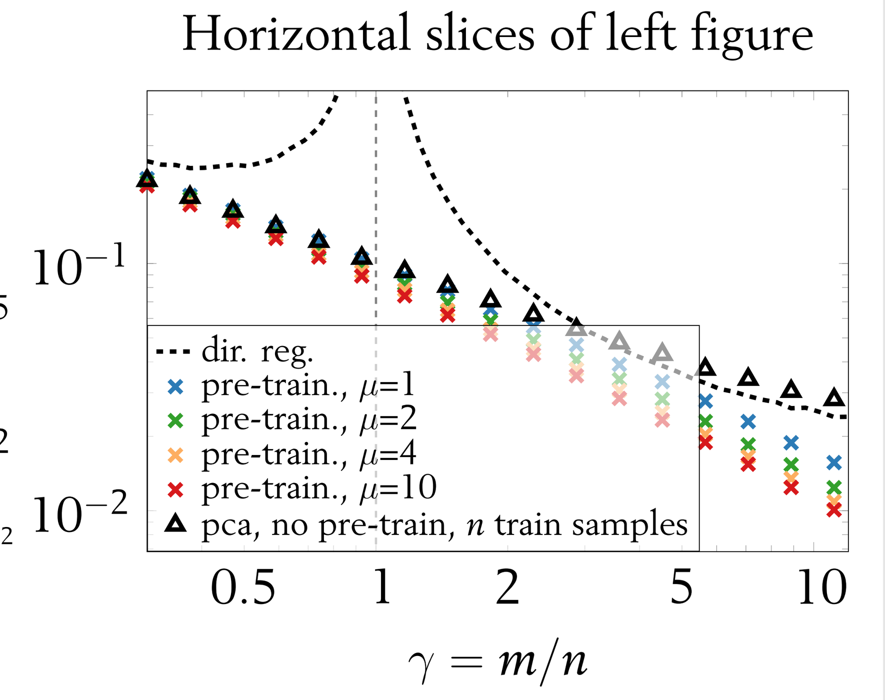
1. Correct latent dimension chosen, i.e. $\hat{d} = d$.
2. Many pre-training samples n_p.
3. Quickly decaying eigenvalues, i.e. $|s_i - s_j|$ large.

Connection to risk

- Xu and Hsu [2] present results for risk with general but known $\hat{V}$.
- Theorem provides missing connection when [2] can be used in practice.

Pre-training – Numerical results





For $\alpha = 0$: more pre-training data n_p does not change risk; horizontal slices equal.

Interpretation

1. Risk decreases for increasing γ; similar to supervised case.
2. $\alpha = 0.25$: risk decreases for more pre-training data n_p; especially for $\gamma > 1$.

→ for $\alpha = 0$ eigenvectors perfectly estimated.
→ for $\alpha = 0.25$ eigenvector estimation improves with more n_p.

Conclusion

Link to paper:

Supervised case:

1. Generalized results from [1] for PCA-regression.
2. Selecting sufficiently large $\hat{d}$ is crucial for low risk.

Pre-training:

1. More pre-training data n_p only help to improve eigenvector estimates.
2. $\alpha > 0$ is necessary such that more pre-training data are helpful.

References

[1] Surprises in highdimensional ridgeless least squares interpolation
Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani
The Annals of Statistics, 50(2):949–986, 2022

[2] On the number of variables to use in principal component regression
Ji Xu and Daniel J Hsu
Advances in neural information processing systems, 32, 2019.

daniel.gedon@it.uu.se