On the number of variables to use in principal component regression

Foundations of Data Science
Data Science Institute

Ji Xu, Daniel Hsu

Computer Science Department and Data Science Institute, Columbia University

Data Science Institute
Columbia University

Principal Component Regression

Model: Suppose the data consists of n i.i.d. observations $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{R}^N \times \mathbb{R}$, where

$$y_i = x_i^\top \theta + w_i,$$

and

$$x_i \sim \mathcal{N}(0, \Sigma), \quad w_i \sim \mathcal{N}(0, \sigma^2).$$

Principal component regression: Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the eigenvalues of Σ. Let $v_1, v_2, \ldots, v_p$ be the corresponding eigenvectors. The PCR estimator $\hat{\theta}$ for θ is defined by (minimum ℓ_2 norm solution for $p > n$ regime)

$$\hat{\theta}_P := \begin{cases} (X_P^\top X_P)^{-1} X_P^\top y & \text{if } p \leq n, \\ X_P^\top (X_P X_P^\top)^{-1} y & \text{if } p > n, \end{cases}$$

where $X_P = [x_1| \cdots |x_n]^\top [v_1| \ldots |v_p]$. The prediction error is given by

$$\text{Error}_p := \mathbb{E}_{x,y}[(y - x^\top \hat{\theta}_P)^2].$$

Question: What is the optimal value of p that minimizes the prediction error?

Double descent phenomenon

- First descent: classic U-shaped risk curve arising from a bias-variance trade-off.
- Second descent: behavior of models in $\mathcal{H}$ that interpolate training data.
- This particular shape is observed in many learning problems, such as neural networks, decision trees and ensemble methods.

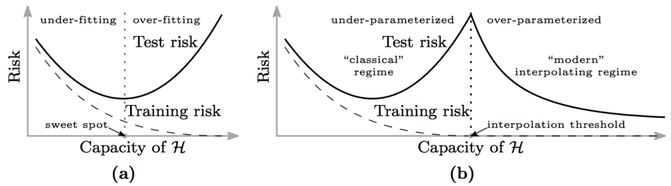

Figure: [BHMM19] (a) The classical U-shaped risk curve arising from the bias-variance trade-off. (b) The double descent risk curve including both the U-shaped risk curve and the observed behavior from using high capacity function classes.

For principal component regression (PCR):

Assumption: We assume $\mathbb{E}_\theta[\theta] = 0$ and $\mathbb{E}_\theta[\theta\theta^\top] = I$.

Question: Does double descent phenomenon happen in PCR?

Question: When the second descent achieve error smaller than the first descent?

Case of Polynomial Decay

We first analyze a special case when the eigenvalues of Σ decay to zero at a polynomial rate. Specifically, we assume

A.1 There exists a constant $\kappa > 0$ such that $\lambda_j = j^{-\kappa}$ for all $j = 1, \ldots, N$.

A.2 There exist constants $\alpha \in [0, 1]$ and $\beta \in (0, 1)$ such that $p/N \to \alpha$ and $n/N \to \beta$ as $p, n, N \to \infty$.

Define $m_\kappa(z)$ for $z \leq 0$ to be the smallest positive solution to the equation

$$-z = \frac{1}{m_\kappa(z)} - \frac{1}{\beta} \int_{\alpha^{-\kappa}}^{\infty} \frac{1}{\kappa t^{1/\kappa} (1 + t \cdot m_\kappa(z))} \, \mathrm{d}t, \tag{1}$$

and let $m'_\kappa(\cdot)$ denote the derivative of $m_\kappa(\cdot)$.

Remark: $m_\kappa(z)$ is the Stieltjes transform of the limiting distribution of the empirical distribution of the eigenvalues of $N^\kappa \Sigma$.

Theorem 1. *Assume* **A.1** *with constant* κ *and* **A.2** *with constants* α *and* β.

(i) *Risk characterization at* $\alpha < \beta$*: For all* $\alpha < \beta$*, we have*

$$\mathbb{E}_{w,\theta}[\text{Error}] \xrightarrow{\text{p}} \left(N^{1-\kappa} \int_\alpha^1 t^{-\kappa} \, \mathrm{d}t + \sigma^2 \right) \cdot \frac{\beta}{\beta - \alpha} =: \mathcal{R}_\kappa(\alpha, \sigma), \quad \forall \alpha < \beta.$$

(ii) *Optimal risk at* $\alpha < \beta$*: When* $\kappa > 1$*, the minimum of* $\mathcal{R}_\kappa(\alpha, \sigma)$ *is achieved at* $\alpha = 0$ *and the minimum risk is given by*

$$\min_{\alpha < \beta} \mathcal{R}_\kappa(\alpha, \sigma) = \sigma^2.$$

When $\kappa \leq 1$*, the minimum of* $\mathcal{R}_\kappa(\alpha, \sigma)$ *is achieved at* α^* *which is the unique solution of the equation* $h_\kappa(\alpha) = 0$ *on* $(0, \beta)$*, where* $h_\kappa(\alpha)$ *is given by*

$$h_\kappa(\alpha) := \frac{\beta}{\alpha} - \int_\alpha^1 t^{\kappa - 2} \, \mathrm{d}t - 1 - \sigma^2 \mathbb{1}\{\kappa = 1\}.$$

The minimum risk is therefore given by

$$\min_{\alpha < \beta} \mathcal{R}_\kappa(\alpha, \sigma) = N^{1-\kappa} \frac{\beta}{(\alpha^*)^\kappa}.$$

(iii) *Risk characterization at* $\alpha > \beta$*: For all* $\alpha > \beta$*, the function* m_κ *defined in* (1) *and its derivative* m'_κ *are well-defined and positive at* $z = 0$*, and*

$$\mathbb{E}_{w,\theta}[\text{Error}] \xrightarrow{\text{p}} N^{1-\kappa} \frac{\beta}{m_\kappa(0)} + \left(N^{1-\kappa} \int_\alpha^1 t^{-\kappa} \, \mathrm{d}t + \sigma^2 \right) \frac{m'_\kappa(0)}{m_\kappa^2(0)} =: \mathcal{R}_\kappa(\alpha, \sigma).$$

(iv) *Comparison between two regimes: When* $\kappa > 1$*, the minimum risk for all* $\alpha < 1$ *and* $\alpha \neq \beta$ *is achieved at* $\alpha = 0$*, i.e.,* $p = o(n)$*. When* $\kappa < 1$*, let* α^* *be the minimizer of* $\mathcal{R}_\kappa(\alpha, \sigma)$ *over the interval* $[0, \beta)$*. Then*

$$\limsup_N \frac{\mathcal{R}_\kappa(1, \sigma)}{\mathcal{R}_\kappa(\alpha^*, \sigma)} < 1.$$

Case of General Decay

Results from Theorem 1 can be extended to the eigenvalues of Σ with other decay rate when the following assumptions hold:

B.1 $\|\Sigma\|_2 \leq C$ for some constant $C > 0$. Also, there exists a positive sequence $(c_N)_{N \geq 1}$ such that the empirical eigenvalue distribution of $c_N \Sigma$ converges as $N \to \infty$ to $F = (1 - \delta) F_0 + \delta F_1$, where $\delta \in (0, 1]$, F_0 is a point mass of 0, and F_1 has a continuous probability density f supported on either $[\eta_1, \eta_2]$ or $[\eta_1, \infty)$ for some constants $\eta_1, \eta_2 > 0$.

B.2 There exist constants $\nu > 0$ and $\beta \in (0, \delta)$ s.t. $p = \sum_{i=1}^N \mathbb{I}(\lambda_i \geq \nu_N)$, $\nu_N c_N \to \nu$ and $n/N \to \beta$ as $n, N \to \infty$.

Remark: **B.1** is the extension of **A.1** with $c_N = N^\kappa$. **B.2** is the extension of **A.2** where ν_N is the threshold parameter that determines the number of selected principal components.

Summaries and Discussions

- We confirm the "double descent" in a natural setting with Gaussian design.
- Optimal p depends on noise level and decay rate of the eigenvalues of Σ.
  - When $\kappa < 1$, a smaller risk is achieved after the interpolation threshold ($p > n$) than any point before ($p < n$).
  - When $\kappa > 1$, a smaller risk is achieved after the interpolation threshold ($p > n$) only in the noiseless setting.
- When Σ is unknown
  - Estimate Σ via unlabeled data.
  - Since the dominance of the $p > n$ regime is always established at $p = N$ (full model), we believe same results hold for standard PCR as well.

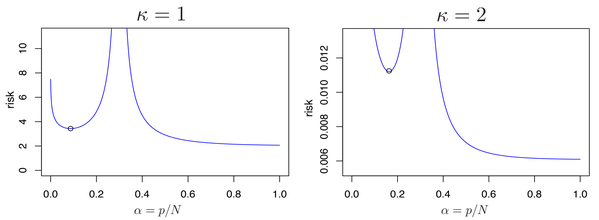

Figure: The asymptotic risk function $\mathcal{R}_\kappa$ as a function of α (with $\sigma = 0$, $n = 300$, $N = 1000$, $\beta = n/N = 0.3$ and $\kappa = 1, 2$ respectively). The location of α^* from Theorem 1 is marked with a black circle. In both cases, the asymptotic risk at $\alpha = 1$ is lower than the asymptotic risk at α^*.

jx2193@columbia.edu, djhsu@cs.columbia.edu