Performance of Empirical Risk Minimization for Principal Component Regression

Christian Brownlees*, Stefán Guðmundsson†, Yaping Wang*

*Universitat Pompeu Fabra and BSE, †Aarhus University and CREATES

Introduction

- Principal component regression (PCR) consists in regressing a prediction target Y_t on the first K principal components extracted from a (large) set of p predictors X_t .
- In this paper we study PCR from a learning theory perspective with dependent data.
- We establish an oracle inequality for PCR. That is, a nonasymptotic prediction performance guarantee for PCR.

Highlights

- **Nonparametric.** The relation between the target and regressors is not specified.
- **Strong and Weak Signals.** The largest K eigenvalues of the covariance matrix of predictors X_t diverge at the rate p^α where
  - $\alpha = 1$ (strong signal case) or
  - $\alpha = (1/2, 1)$ (weak signal case).
- **Focus on Predictive Performance.** We study the gap between the predictive performance of PCR and the best linear predictor.

Principal Component Regression

- $\{(Y_t, \boldsymbol{X}_t)'\}_{t=1}^T$: Stationary sequence of dependent random vectors taking values in $\mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^p$
  - The relation between Y_t and $\boldsymbol{X}_t$ is not specified
  - The number of predictors p is non negligible relative to T
- Predicting Y_t using $\boldsymbol{X}_t$: rely on principal component regression
  1. Compute the first K principal components of $\boldsymbol{X}_t$, $\hat{\boldsymbol{P}}_t$ is given by (s.t. normalization)
  $$\hat{\boldsymbol{P}}_t = \hat{\boldsymbol{\Lambda}}_K^{-1/2} \hat{\boldsymbol{V}}_K' \boldsymbol{X}_t,$$
  where $\hat{\boldsymbol{\Lambda}}_K$ is the $K \times K$ diagonal matrix consists of the largest K eigenvalue of $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}/T$ and $\hat{\boldsymbol{V}}_K$ is the $p \times K$ matrix of corresponding eigenvectors
  2. Run a regression of Y_t on the principal components $\hat{\boldsymbol{P}}_t$ to obtain the K-dimensional coefficients
  $$\hat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \frac{1}{T} \sum_{t=1}^T (Y_t - \boldsymbol{\vartheta}' \hat{\boldsymbol{P}}_t)^2$$

Performance of PCR

- PCR is in fact a procedure of choosing a prediction rule by minimizing the empirical risk
- Predictive performance of PCR is measured by the conditional risk:
  $R(\hat{\boldsymbol{\vartheta}}) = \mathbb{E}\left[(Y_t - \hat{\boldsymbol{\vartheta}}' \hat{\boldsymbol{P}}_t)^2 \middle| \mathcal{D}\right]$, $\mathcal{D}$:data used for estimation
- The natural benchmark is the risk of the best linear predictor
  $$R(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} R(\boldsymbol{\theta}), \quad \text{where } R(\boldsymbol{\theta}) = \mathbb{E}\left[(Y_t - \boldsymbol{X}_t'\boldsymbol{\theta})^2\right] .$$
- Oracle inequality:
  $$R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) \leq B_T(K, p)$$
  holds at least with probability $1 - \delta_T(p, K)$ for all sufficiently large T, where $B_T(p, K)$ and $\delta_T(p, K)$ approach 0 as $T \to \infty$.
  - We care about prediction accuracy, not about estimation accuracy.
  - We care about achieving optimality relative to the class of forecasting rules (here linear forecasts), not relative to the "true model"
  - We care about finite sample guarantees, not about asymptotic ones.

Optimal Learning Rate

Key question

We care about the rate of convergence $B_T(p, K)$. What is the optimal learning rate $B_T(p, K)$ that can be achieved?

- This in general can be a tough question to answer. In this paper we shed partial light onto this question by comparing the learning rates obtained in this work with the optimal learning rate that could be achieved if the principal components were observed.
- It is well known that in such a case the optimal rate of convergence for linear aggregation is of the order K/T, which is achieved by the least squares estimator [Tsybakov 2003]
- Challenge: The principal components are not not observed and have to be estimated from the data

Proof Strategy

The proof is based on a decomposition (we have omitted the rotation matrix of the principal components):

$$\begin{aligned} R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) &= \|Y_t - \hat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 \\ &+ \|Y_t - \boldsymbol{P}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 \\ &+ \|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 \\ &+ \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 - \|Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^*\|_{L_2}^2 \\ &= A_T + B_T + C_T + D_T \ , \end{aligned}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\vartheta}} &= \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \frac{1}{T} \sum_{t=1}^T (Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta})^2, \\ \boldsymbol{\vartheta}^* &= \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \mathbb{E}[(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta})^2] \\ \boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}[(Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^*)^2] \ . \end{aligned}$$

Notation: For a generic random variable X, $\|X\|_{L_2}$ is defined as $[\mathbb{E}|X|^2]^{\frac{1}{2}}$

- A_T and B_T capture the risk of PCR that is due to the estimation of the principal components.
- C_T captures the risk of the least squares estimator of PCR based on the population principal components and it achieves the optimal rate K/T (We rely on a proof strategy based on the so-called small-ball method [Mendelson 2015, Lecué and Mendelson 2016]).
- D_T is a constant representing (squared) bias or approximation error.

Main Result

Theorem: Performance of Empirical Risk Minimization

Suppose that all assumptions are satisfied. Then for any $\eta > 0$ there exists a constant $C > 0$ such that, for any T sufficiently large,

$$R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) \leq 2(\boldsymbol{\theta}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\theta}^* + C\left[\frac{1}{p^{2\alpha-1}} + \left(\frac{p}{Tp^\alpha}\right)^2 p^{\frac{2}{r_\alpha}} + \frac{K}{T}\right]\log(T),$$

holds with probability at least $1 - T^{-\eta}$, where $\boldsymbol{V}_R$ and $\boldsymbol{\Lambda}_R$ are the matrices of eigenvectors and eigenvalues corresponding to the $K+1$ to p eigenvalues of $\boldsymbol{\Sigma}$.

- The bound is made up of two terms. The first can interpreted as the approximation error of PCR and the second as the estimation error.
- For ease of exposition we assume that the approximation error is at most of the same order of magnitude of the estimation error:
  - The optimal rate can be achieved (up to a logarithmic factor) in the both the strong signal ($\alpha = 1$), and the weak signal ($\alpha < 1$) cases, provided that $\alpha > 2/3$.
  - The larger the values of α, r_α and r_K, the larger the range of admissible growth rates for the number of predictors r_p.

Conclusion

- We establish prediction performance guarantees for empirical risk minimization for principal component regression.
- Analysis is carried out in a nonparametric framework. In particular the target variable Y_t is not assumed to be generated by a factor model.
- Under appropriate conditions, PCR achieves optimal performance (up to a logarithmic factor) in both the strong signal and weak signal regimes.

Contact: yaping.wang@barcelonagse.eu