Choi, Taylor and Tibshirani (2017, AoS)

Prelim. for Matsumura & Tachibana (forthcoming)

Yasuyuki Matsumura

Last Updated: August 28, 2025

yasu0704xx.github.io

The Annals of Statistics
2017, Vol. 45, No. 6, 2590–2617
https://doi.org/10.1214/16-AOS1536

SELECTING THE NUMBER OF PRINCIPAL COMPONENTS: ESTIMATION OF THE TRUE RANK OF A NOISY MATRIX

BY YUNJIN CHOI*, JONATHAN TAYLOR†,1 AND ROBERT TIBSHIRANI†,2

*National University of Singapore** *and Stanford University*†

Principal component analysis (PCA) is a well-known tool in multivariate statistics. One significant challenge in using PCA is the choice of the number of principal components. In order to address this challenge, we propose distribution-based methods with *exact* type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples. Assuming Gaussian noise, we derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of [Taylor, Loftus and Tibshirani (2013)] in a PCA setting. In simulation studies, we find that our proposed methods compare well to existing approaches.

Choi, Taylor, Tibshirani (2017, AoS)

- In using principal component analysis (PCA), the choice of the number of principal components is a significant challenge.
- Choi, Taylor and Tibshirani (2017) propose distribution-based methods with exact type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples.
- Assuming Gaussian noise, Choi, Taylor and Tibshirani (2017) derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of Taylor, Loftus and Tibshirani (2016) in a PCA setting.

Contents

Introduction

Proposed Distribution-Based Methods

Rank Estimation

Estimationg the Noise Level

Additional Examples

Conclusions