# Permutation Tests for Identifying Number of Factors for High-Dimensional Time Series


THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Sixing Hao

Supervisor: Qiwei Yao

Department of Statistics, London School of Economics and Political Science


## Introduction

**Motivation**: Dimension reduction on high-dimensional time series, through revealing its underlying process.

**Background**:

- This project utilizes the idea of *Factor Modeling* for high-dimensional time series.
- By *Factor Modeling*, a $p$-dimensional time series $\mathbf{Y}$ is composed of linear mixture of $r$-dimensional time series $\mathbf{X}$, where $r < p$.

**Goal**: Estimation on number of factors $r$.

**Contribution**: An estimation method on $r$ when factors have different strength.

## Factor Model

$$\underset{(n\times p)}{\mathbf{Y}} = \underset{(n\times r)}{\mathbf{X}}\,\underset{(r\times p)}{\mathbf{A}^{\top}} + \underset{(n\times p)}{\epsilon} \tag{1}$$

- $\mathbf{X}$ is an unobserved latent process with $r \le p$.
- $\mathbf{A}$ is a $p \times r$ constant factor loading matrix with rank $r$.
- $\epsilon$ is a vector white-noise process.

**Factor Strength**

Let $a \asymp b$ if $a = O(b)$ and $b = O(a)$, assume that for $\mathbf{A} = (a_1, \dots, a_r)$,

$$\|\mathbf{a}_j\|_2^2 \asymp p^{1-\delta_j}, \quad j = 1, \dots, r, \quad \delta_j \in [0, 1],$$

If $\delta_j = 0$, $\mathbf{X}_j$ is called a **strong factor**. Else, $\mathbf{X}_j$ is called a **weak factor**.

## Permutation Tests

Permutation test checks for serial correlation of a time series. The hypothesis is

$$H_0: \forall k \in [1, m],\ \rho(k) = 0, \quad H_A: \exists k \in [1, m]\, s.t.\ \rho(k) \neq 0,$$

where $\rho(k)$ represents the autocorrelation at lag $k$, and $m$ is the maximal lag to be considered.

**Steps for Permutation Tests**

1. Given a time series $S_{obs} = s_1, \dots, s_n$, permute its elements to get $S_\pi = s_{\pi(1)}, \dots, s_{\pi(n)}$, Repeat $L$ times.
2. Choose test statistic for serial correlation: $T(\cdot) = n(n+2)\sum_{k=1}^{m} \frac{m-k+1}{m}\frac{\hat{\rho}_k^2}{n-k}$, calculate $p$-value by:

$$p\text{-}value = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}(T(S_{\pi_i}) \ge T(S_{obs})). \tag{2}$$

3. If $p$-$value \le \alpha$, $S_{obs}$ is **serially correlated**. $\alpha$ is the pre-specified significance level.

## Estimation on Number of Factors

1. Use covariance matrix of $\mathbf{Y}$ at lag $k$: $\mathbf{\Sigma}_y(k) = Cov(y_{t+k}, y_t)$, gather information across multiple lags by

$$\mathbf{M} = \sum_{k=1}^{m} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)^{\top}, \quad m \ge 1. \tag{3}$$

2. Perform eigen-decomposition on $\mathbf{M}$. Define $\mathbf{\Gamma} = (b_1, \dots, b_p)$, where $(b_1, \dots, b_p)$ are eigenvectors of $\mathbf{M}$ in descending order of corresponding eigenvalues. $\mathbf{\Gamma}$'s columns contain estimation of $\mathbf{A}$ and noise.
3. Define $\mathbf{Z} = \mathbf{Y\Gamma}$, which is an approximation of $\mathbf{X}$. Conduct permutation tests on all $p$ columns of $\mathbf{Z}$, obtain a sequence of p-values $(p_1, \dots, p_p)$.
4. Obtain the estimator by identifying number of columns with significant serial correlation:

$$\hat{r}_{PT} = \sum_{i=1}^{p} \mathbb{1}(p_i \le \alpha). \tag{4}$$

## Estimation Accuracy of $\hat{r}_{PT}$

We demonstrate our estimator through 2 sets of simulations. In first setting, all factors are strong ($\delta_i = 0$). In second setting, $\delta_i \sim Unif(0, 1)$. For both simulations,

- $n = (400, 900, 1600, 2500, 3600)$, $p = (\sqrt{n}, 0.5n, n, 2n)$, $r = 9$, $L = 1000$.
- Factor within $\mathbf{X}$ from AR(1) process.
- Elements of $\epsilon \sim N(0, 1)$, independent of time, elements of $\mathbf{A} \sim N(0, 1)$.
- Maximal lag for covariance estimation is 1.

To compare, $\hat{r}_{Ratio}$ from Lam&Yao (2012) is added, which is based on ratio of eigenvalues of $\mathbf{M}$.

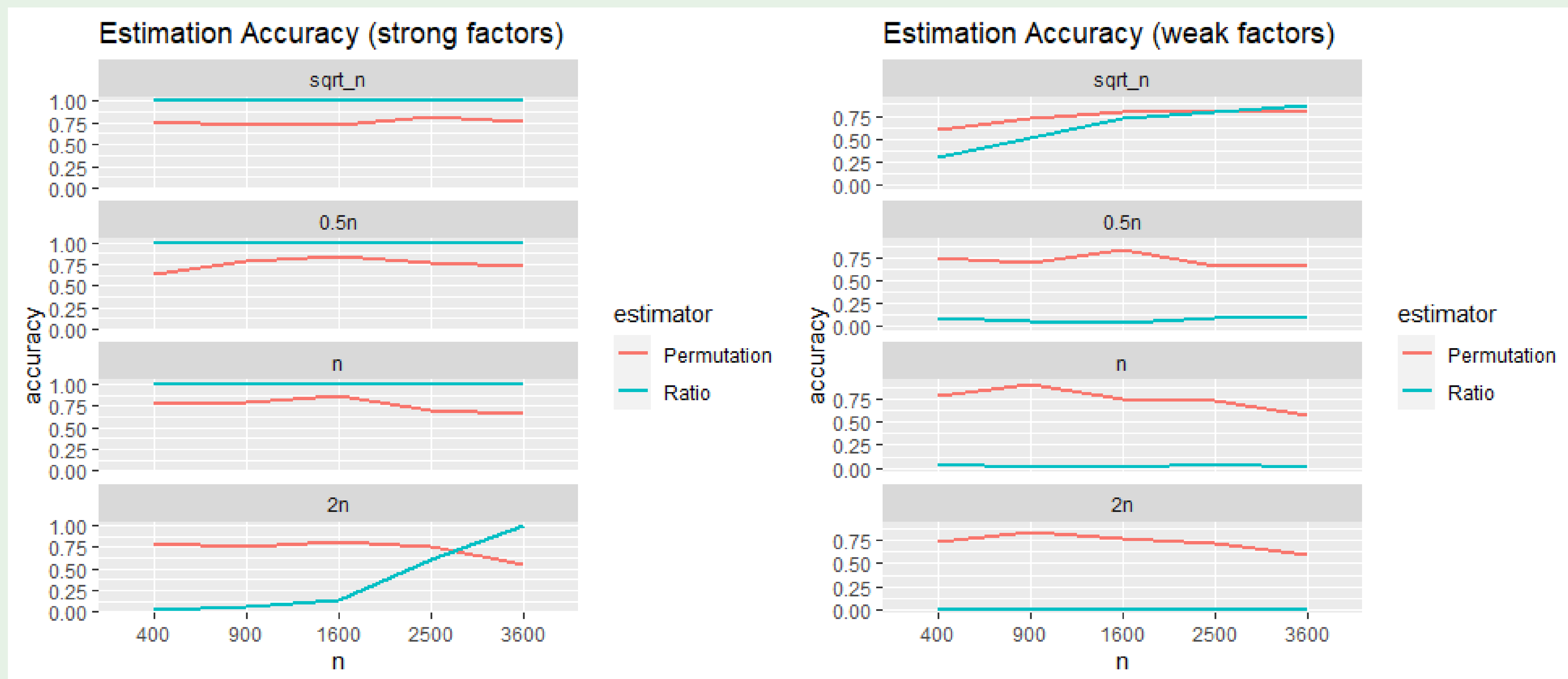


Figure: Accuracy of estimators from Permutation Test and Ratio Test with 100 repetitions. Figure on left represents the result when all factors are strong, and figure on right is when all factors are weak with different strength.

**Remark:** $\hat{r}_{Ratio}$ has better performance when factors are strong and dimension $p < 2n$. If factors are weak, $\hat{r}_{PT}$ is stable at a reasonable accuracy level, yet $\hat{r}_{Ratio}$ cannot give good estimation.

## Mean and Standard Deviation of $\hat{r}_{PT}$ and $\hat{r}_{Ratio}$

| | n | p | r0 | PT_mean | PT_sd | Ratio_mean | Ratio_sd |
|---|---|---|---|---|---|---|---|
| 1 | 400.00 | 20.00 | 9.00 | 8.51 | 1.60 | 9.00 | 0.00 |
| 2 | 400.00 | 200.00 | 9.00 | 8.58 | 1.81 | 9.00 | 0.00 |
| 3 | 400.00 | 400.00 | 9.00 | 8.44 | 1.53 | 9.00 | 0.00 |
| 4 | 400.00 | 800.00 | 9.00 | 8.52 | 1.21 | 386.33 | 66.69 |
| 5 | 900.00 | 30.00 | 9.00 | 8.83 | 1.54 | 9.00 | 0.00 |
| 6 | 900.00 | 450.00 | 9.00 | 8.82 | 1.32 | 9.00 | 0.00 |
| 7 | 900.00 | 900.00 | 9.00 | 8.65 | 1.08 | 9.00 | 0.00 |
| 8 | 900.00 | 1800.00 | 9.00 | 8.45 | 1.49 | 844.66 | 212.19 |
| 9 | 1600.00 | 40.00 | 9.00 | 9.16 | 1.25 | 9.00 | 0.00 |
| 10 | 1600.00 | 800.00 | 9.00 | 9.01 | 0.87 | 9.00 | 0.00 |
| 11 | 1600.00 | 1600.00 | 9.00 | 8.99 | 0.73 | 9.00 | 0.00 |
| 12 | 1600.00 | 3200.00 | 9.00 | 8.91 | 1.07 | 1375.54 | 554.14 |
| 13 | 2500.00 | 50.00 | 9.00 | 9.27 | 0.68 | 9.00 | 0.00 |
| 14 | 2500.00 | 1250.00 | 9.00 | 9.12 | 0.94 | 9.00 | 0.00 |
| 15 | 2500.00 | 2500.00 | 9.00 | 9.08 | 1.08 | 9.00 | 0.00 |
| 16 | 2500.00 | 5000.00 | 9.00 | 9.12 | 0.94 | 1004.60 | 1225.50 |
| 17 | 3600.00 | 60.00 | 9.00 | 9.20 | 0.74 | 9.00 | 0.00 |
| 18 | 3600.00 | 1800.00 | 9.00 | 9.19 | 1.08 | 9.00 | 0.00 |
| 19 | 3600.00 | 3600.00 | 9.00 | 9.22 | 1.14 | 9.00 | 0.00 |
| 20 | 3600.00 | 7200.00 | 9.00 | 9.42 | 1.26 | 9.00 | 0.00 |

(a) Mean & SD of $\hat{r}_{PT}$ and $\hat{r}_{Ratio}$ for Strong Factor Setting

| | n | p | r0 | PT_mean | PT_sd | Ratio_mean | Ratio_sd |
|---|---|---|---|---|---|---|---|
| 1 | 400.00 | 20.00 | 9.00 | 8.02 | 2.08 | 5.70 | 3.21 |
| 2 | 400.00 | 200.00 | 9.00 | 8.76 | 1.66 | 4.17 | 2.50 |
| 3 | 400.00 | 400.00 | 9.00 | 8.90 | 1.44 | 3.78 | 2.34 |
| 4 | 400.00 | 800.00 | 9.00 | 8.56 | 1.81 | 398.62 | 0.49 |
| 5 | 900.00 | 30.00 | 9.00 | 9.05 | 1.20 | 7.02 | 2.74 |
| 6 | 900.00 | 450.00 | 9.00 | 9.07 | 1.41 | 4.01 | 2.45 |
| 7 | 900.00 | 900.00 | 9.00 | 8.82 | 1.30 | 4.11 | 2.11 |
| 8 | 900.00 | 1800.00 | 9.00 | 9.03 | 0.83 | 898.00 | 0.00 |
| 9 | 1600.00 | 40.00 | 9.00 | 9.22 | 0.72 | 7.64 | 2.68 |
| 10 | 1600.00 | 800.00 | 9.00 | 9.18 | 0.56 | 4.09 | 2.32 |
| 11 | 1600.00 | 1600.00 | 9.00 | 9.10 | 1.18 | 4.42 | 1.89 |
| 12 | 1600.00 | 3200.00 | 9.00 | 9.20 | 0.75 | 1598.00 | 0.00 |
| 13 | 2500.00 | 50.00 | 9.00 | 9.18 | 1.04 | 7.79 | 2.69 |
| 14 | 2500.00 | 1250.00 | 9.00 | 9.37 | 0.86 | 4.49 | 2.44 |
| 15 | 2500.00 | 2500.00 | 9.00 | 9.22 | 1.07 | 4.28 | 1.94 |
| 16 | 2500.00 | 5000.00 | 9.00 | 9.28 | 0.94 | 2498.00 | 0.00 |
| 17 | 3600.00 | 60.00 | 9.00 | 9.04 | 0.94 | 7.97 | 2.63 |
| 18 | 3600.00 | 1800.00 | 9.00 | 9.51 | 0.87 | 4.70 | 2.36 |
| 19 | 3600.00 | 3600.00 | 9.00 | 9.50 | 0.64 | 4.45 | 1.86 |
| 20 | 3600.00 | 7200.00 | 9.00 | 9.59 | 1.22 | 3598.00 | 0.00 |

(b) Mean & SD of $\hat{r}_{PT}$ and $\hat{r}_{Ratio}$ for Weak Factor Setting

**Remark:** $\hat{r}_{PT}$ have more stable standard deviation across different settings, while $\hat{r}_{Ratio}$ has perfect performance only when factor strength is strong and $p < 2n$ for small $n$.

## Conclusions

**Performance of $\hat{r}_{PT}$**

- Estimation accuracy of $\hat{r}_{PT}$ is generally reasonable.
- Inaccurate estimations $\hat{r}_{PTwrong} = r \pm 1$ most of the time.
- $\hat{r}_{PT}$ has similar performance in both settings.
- Accuracy drops when both $n$ and $p$ are large, due to estimation error of sample covariance matrix under large sample size.

**Comparing $\hat{r}_{PT}$ with $\hat{r}_{Ratio}$**

- When factors are **strong**, $\hat{r}_{PT}$ is better than $\hat{r}_{Ratio}$, only if $p$ is relative large.
- $\hat{r}_{PT}$ can handle factors with different levels of strength, yet $\hat{r}_{Ratio}$ works only when all factors have same strength.
- $\hat{r}_{PT}$ can perform well with small $n$, even when $p$ is large.

## Future Work

1. Better estimation of **sample covariance matrix** at large sample size.
2. Choice of **Test Statistics** might need adjustment for deriving asymptotic properties.
3. Improvement on speed of estimation is needed.
4. Information from ordering of eigenvalues can be further utilized by designing new estimators.

## Key References


- Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 694-726.
- Romano, J. P., & Tirlea, M. A. (2022). Permutation testing for dependence in time series. *Journal of Time Series Analysis*, 43(5), 781-807.
- Fisher, T. J., & Gallagher, C. M. (2012). New weighted portmanteau statistics for time series goodness of fit testing. Journal of the American Statistical Association, 107(498), 777-787.