Dimension Selection in High-dimensional Dataset

Introducing five different methods of selecting the optimal dimensions in high-dimensional data setting and comparing performances of each method

MP Law

Marchenko-Pastur Law states the average distribution of eigenvalues is bounded when the column dimensions of correlation matrix ***p*** →∞ and its aspect ratio with sample size ***n*** converges to a constant, where ***n/p* → *c* ∈ *(0, ∞)*.** The bounded spectral distribution is as follows:

$$\mathbf{f_c(y) = \frac{1}{2\pi y}\sqrt{(y - c_-)(c_+ - y)}}$$

and the edge points c± are derived as :

$$\mathbf{c_- = \left(1 - \sqrt{\tfrac{1}{c}}\right)^2} \quad \text{and} \quad \mathbf{c_+ = \left(1 + \sqrt{\tfrac{1}{c}}\right)^2}$$

Since all the eigenvalues are non-zero, the selection criterion is retaining all the eigenvalues greater than c+. MP Law performs the best especially when ***p >> n.***

EKC

Empirical Kaiser Criterion considers both Kaiser's rule and Marchenko-Pastur Law. It also takes into account the serial nature of eigenvalues. The cut-off value for each eigenvalue is denoted as:

$$\theta_j^{EKC} = max\left[\frac{p - \sum_{j=0}^{j-1}\theta_j}{p - j + 1}\left(1 + \sqrt{\frac{1}{c}}\right)^2, \ 1\right], with \ \theta_0 = 0$$

Eigenvalues greater than its respect cut-off value is retained as selected latent factors. EKC is a competitive candidate to Marchenko-Pastur Law, especially when ***p*** is moderate to ***n*** and factor loadings are high.

Eigenvalue Ratio

Eigenvalue ratio method takes the largest ratio of two adjacent eigenvalues from correlation matrix. The index of the largest eigenvalue ratio is the optimal number of latent factors. This method is easy-to-apply, but overestimates the number of latent factors when ***p > n***. Eigenvalue ratio denotes as:

$$K_{ER} = \underset{i \leqslant k_{max}}{argmax}\frac{\theta_i}{\theta_{i+1}}$$

Growth Ratio

Growth ratio method takes into account the average squared residuals from regression on each factors. The growth ratio is calculated as follows:

$$G_{GR} = \underset{i \leqslant g_{max}}{argmax}\frac{\ln\left[V(i-1)/V(i)\right]}{\ln\left[V(i)/V(i+1)\right]} = \frac{\ln(1+\theta_i^*)}{\ln(1+\theta_{i+1}^*)}$$

The numerator and denominator of the growth ratio equation both represent the growth rates of squared residuals from accounting for one less latent factors in the regressions. The index of the largest growth ratio is the optimal number of latent factors. Growth ratio method shares the similar rationale as eigenvalue ratio method and performs slightly better than it.

Kaiser's rule

Kaiser's rule states that the optimal number of latent factors are selected as the number of eigenvalues greater than 1. The easiness of this method attracts a wide range of users, but its estimating accuracy cannot be guaranteed in most of the cases, especially when ***p > n***.