UNIVERSITY OF CALIFORNIA
Los Angeles

Empirical Test of Applicability of Donoho and Gavish's Method in Determining the Number of Factors in Factor Analysis

A thesis submitted in partial satisfaction
of the requirements for the degree Master of Science
in Statistics

By

Yuxi Zhou

2014

ABSTRACT OF THE THESIS

Empirical Test of Applicability of Donoho and Gavish's Method in Determining the Number of Factors in Factor Analysis

By

Yuxi Zhou

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Yingnian Wu, Chair

Donoho and Gavish (2013) proposed a method of recovering a matrix by selecting singular values above a hard threshold. In their paper, $4/\sqrt{3}$ is proved to be asymptotic MSE-optimal choice of hard threshold. We empirically test the applicability of Donoho and Gavish's method in factor analysis with simulated datasets and assess its asymptotic property when both the matrix and sample size grow while keeping the true number of factors fixed.

The thesis of Yuxi Zhou is approved.

Peter M. Bentler

Frederic R. Paik Schoenberg

Nicolas Christou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2014

Table of Contents

Acknowledgements

Deepest appreciation to my advisor, Professor Peter M. Bentler. Without his persistent help and encourage, this thesis wouldn't have been possible.

Empirical Test of Applicability of Donoho and Gavish's Method in Determining the Number of Factors in Factor Analysis

Abstract

Donoho and Gavish (2013) proposed a method of recovering a matrix by selecting singular values above a hard threshold. In their paper, $4/\sqrt{3}$ is proved to be asymptotic MSE-optimal choice of hard threshold. We empirically test the applicability of Donoho and Gavish's method in factor analysis with simulated datasets and assess its asymptotic property when both the matrix and sample size grow while keeping the true number of factors fixed.

Introduction

Factor analysis has been a useful tool in many fields such as psychology and economics and the selection of number of factors in factor analysis has been a critical issue. There have been a few criteria for determining the number of factors to retain. The Eigenvalue-greater-than-one rule (Guttman, 1954; Kaiser, 1960, 1970)) suggests that eigenvectors with eigenvalues of the correlation matrix greater than 1 should be used to represent the number of factors. It is simply saying that factor analysis should extract factors whose eigenvalue is greater than average (Nunnally & Bernstein, 1994). However, this approach tends to severely overestimate the number of components (Awick & Velicer 1986). Horn (1965) proposed Parallel Analysis, which generates a large number of random datasets with the same pattern of the real dataset for determining which number of factors is most appropriate. Velicer (1976) proposed Minimum Average Partial (MAP), which was shown to underestimate the true number of factors (Hayton et al., 2004). Zwick and Velicer (1986) compared the four methods with simulated datasets and found Parallel Analysis and MAP are more accurate than the other two.

Selecting the number of factors is also an important issue in Bayesian Factor Analysis. The AIC (Akaike, 1973) and BIC (Schwarz, 1978) information criteria have been widely used, and some other selection procedures like Bozdogan (1987) and Shegemasu (1999) have also been developed. Hirose et al. (2011) proposed the GBIC method, which avoids improper solutions caused by the maximum likelihood estimation and was found to be more accurate than BIC (Hirose et al., 2011).

Donoho and Gavish's (2013) study assumes X is an m by n matrix , whose rank is relatively small to its size, Y is the observed noisy m-by-n matrix, and Y = X + σZ, where σ is a scalar and Z has independent, identically distributed entries with zero mean and unit variance. They proposed that the recovery of X can be achieved by selecting singular values above a hard threshold. In their paper, $(4/\sqrt{3})\sqrt{n}\sigma$ is proved to be asymptotic MSE-optimal choice of hard threshold when σ is known.

Donoho and Gavish's method can be applied to the recovery of covariance matrix and correlation matrix when the assumptions are met. However, in the situation of factor analysis when the covariance matrix Σ is based on equal unique variances, thus meeting the DG conditions, the assumption no longer holds in the correlation matrix $\mathrm{P} = \mathrm{diag}(\Sigma)^{-1/2}(\Sigma)\mathrm{diag}(\Sigma)^{-1/2}$. DG's method is probably not applicable to factor analysis and data transformation or other methods should be proposed to meet the conditions.

In the paper, we apply DG's method to determining the number of factors in factor analysis in the special case of equal unique variances that should meet the DG conditions, and compare this approach with another hard threshold method, the eigenvalue-greater- -than-one rule, using simulated datasets.

Design

For the simulations, we start with Gaussian data matrix generated from a given covariance matrix Σ, with loading structure $\Sigma = \Lambda\Phi\Lambda' + \Psi^2$,

$$\Lambda' = \begin{bmatrix} 1.0 & 1.0 & 1.0 & & & & & & & & & & & & \\ & & & 1.0 & 1.0 & 1.0 & & & & & & & & & \\ & & & & & & 1.0 & 1.0 & 1.0 & & & & & & \\ & & & & & & & & & 1.0 & 1.0 & 1.0 & & & \\ & & & & & & & & & & & & 1.0 & 1.0 & 1.0 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} 1.0 & .5 & .5 & .5 & .5 \\ .5 & 1.0 & .5 & .5 & .5 \\ .5 & .5 & 1.0 & .5 & .5 \\ .5 & .5 & .5 & 1.0 & .5 \\ .5 & .5 & .5 & .5 & 1.0 \end{bmatrix}, \text{ and } \Psi^2 = I$$

Λ is a factor loading matrix and Ψ^2 represents the covariance matrix with errors with equal unit variance. Σ represents the covariance structure of a model with 15 variable and 5 factors. Therefore, the sample covariance matrix will be a 15*15 matrix and the X matrix to be recovered is of rank 5.

In order to simulate the situation of growing matrix size and sample size, we then generate datasets with number of variables p=15, 30, 45, sample size N=100, 300, 1000, 5000, respectively. In the case of 30 variables, the factor loading matrix Λ equals $\Lambda_{5\times30}$, and in the case of 45, the factor loading equals $\Lambda_{5\times45}$,

$$\Lambda'_{5\times30} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & & & \cdots & & & & & & & & \\ & & & & & & 1.0 & 1.0 & & & & & & & & & \\ \vdots & & & & & & & & \ddots & & & & & & & & \vdots \\ & & & & & & & & & 1.0 & 1.0 & & & & & & \\ & & & & & & & & \cdots & & & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}$$

$$\Lambda'_{5\times45} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & \cdots & & & & & & & & & \\ \vdots & & & & & & & & & \ddots & & & & & & & & & \vdots \\ & & & & & & & & & \cdots & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}$$

We then use three methods to determine the number of factors: DG's method on the covariance matrix, DG's method on the correlation matrix, and the eigenvalue--greater-than-one criterion on the correlation matrix.

Results

Simulated datasets with p variables and N cases are repeatedly generated 500 times, then mean, median, standard deviation of the number of selected singular values across the 500 replications are included in the following tables.

Table 1. Number of retained singular values when N=100

N=100			
		P=15	
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	3.718	1	4.528
Median	4	1	5
SD	0.568	0	0.519
		P=30	
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	5.012	2.37	5.664
Median	5	2	6
SD	0.109	0.598	0.651
		P=45	
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	5.982	4.378	8.184
Median	6	4	8
SD	0.621	0.544	1.010

When p=15, DG’s on correlation matrix invariably selects 1 singular value and the average number becomes 4.378 when p increases to 45. The other two methods are more accurate when p=15 and p=30. However, when p=45, DG’s on covariance matrix selects on average 5.982 singular values and Eigenvalue-greater--than-one selects on average 8.184, which means that this approach overestimates the number of factors.

Table 2. Number of retained singular values when N=300

N=300			
		P=15	
	DG on Cov. Matrix	DG on Cor. Matrix	e.v>1
Mean	3.916	1	4.94
Median	4	1	5
SD	0.571	0	0.246
		P=30	
	DG on Cov. Matrix	DG on Cor. Matrix	e.v>1
Mean	5	1.668	5
Median	5	2	5
SD	0	0.561	0
		P=45	
	DG on Cov. Matrix	DG on Cor. Matrix	e.v>1
Mean	5	4.834	5.004
Median	5	5	5
SD	0	0.383	0.063

The second table where N=300 has a similar pattern as Table 1 when p=15 and 30, when the DG's method on the covariance matrix and Eigenvalue-greater--than-one are fairly accurate and DG's method on the correlation matrix greatly underestimates the number of factors. However, when p=45 all three methods work quite well, with DG on the covariance matrix best overall.

Table 3. Number of retained singular values when N=1000

N=1000			
		P=15	
	DG on Cov. Matrix	DG on Cor. Matrix	e.v>1
Mean	4.436	1	5
Median	4	1	5
SD	0.564	0	0
		P=30	
	DG on Cov. Matrix	DG on Cor. Matrix	e.v>1
Mean	5	1.038	5
Median	5	1	5
SD	0	0.191	0

	P=45		
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	5	4.994	5
Median	5	5	5
SD	0	0.077	0

In the third table of N=1000, when p=15, DG's on correlation matrix invariably selects 1 singular value while the number gets to around 5 when p increases to 45, which suggests that it has good asymptotic property. The other two methods outperform it when p=15 and p=30 and select exactly 5 factors when p=45.

Table 4. Number of retained singular values when N=5000

N=5000			
	P=15		
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	4.997	1	5
Median	5	1	5
SD	0.045	0	0
	P=30		
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	5	1	5
Median	5	1	5
SD	0	0	0
	P=45		
	DG on Cov. Matrix	**DG on Cor. Matrix**	**e.v>1**
Mean	5	5	5
Median	5	5	5
SD	0	0	0

In the fourth table for the case of N=5000, when p=15, DG's on correlation matrix still invariably selects 1 singular value and it accurately selects 5 when p increases to 45. This empirically proves its asymptotic property, though this is unexpected since the correlation matrix does not meet the DG conditions (Donoho& Gavish, 2013). The other two methods that select 5 factors, are accurate as well.

Since each dataset was generated 500 times within a condition, we also collect the frequencies of correct solutions in which exactly 5 factors are selected. The frequency tables are as follows, showing that DG on the covariance matrix is best overall when p is large (30 or 45), while the Eigenvalue greater than one rule is best with the smallest number of variables.

Table 5 Frequency table of 500 simulations for DG's on covariance matrix

	p=15	P=30	P=45
N=100	30	493	88
N=300	62	500	500
N=1000	236	500	500
N=5000	499	500	500

Table 6 Frequency table of 500 simulations for DG's on correlation matrix

	p=15	P=30	P=45
N=100	0	0	240
N=300	0	0	419
N=1000	0	0	497
N=5000	0	0	500

Table 7 Frequency table of 500 simulations for Eigenvalue-greater-than-one rule

	p=15	P=30	P=45
N=100	298	211	0
N=300	471	500	498
N=1000	500	500	500
N=5000	500	500	500

Discussion

The results from the simulations show that when we keep number of variables fixed at p=15, DG's covariance based method selects somewhat less than 5 factors and selects exactly 5 when the sample size gets larger. However, the DG correlation matrix method it always selects 1 factor unless p and the sample size are very large. One reason that DG on the correlation matrix does not work well is that the equal unique variance assumption is violated. The Eigenvalue- greater-than-one criterion generally works well although it is outperformed by GD on the covariance matrix with the larger number of variables.

Another way to look at the results is to fix the sample size at look at the asymptotic properties of the method, since the GD methodology is based on the idea that both p and N are large. When p is 30 or more, and N is 300 or more, the DG covariance based method performs perfectly. Therefore, when p and N are large enough, and its assumptions are met, GD's method is asymptotically optimal and performs correctly for determining the number of factors. When GD's assumptions are not met, as in its application to the correlation matrix based on a structure with unequal unique variances, the approach generally fails.

There are several limitations regarding this study. First, only one factor loading structure is used, which might not be representative enough. We need to test with samples generated from different population factor loading structures. Second, we set X to be a size 15 by 15, rank 5 matrix, up to one that is 45 by 45 of rank 5, whose size might not be relatively large enough compared to its rank. It is possible that changing the p:rank ratio even better performance of DG's method could be achieved.

Finally, the covariance structure studied here had equal unique variances, which is a condition unlikely to occur in practice. To deal with the unequal unique variance problem, there are three possible directions that might be considered

(1) Transformation on observed data matrix: Suppose Y is the observed data matrix, let $Y' = AY$ be a transformation from Y to meet the condition that correlation matrix of Y' has the structure of $X + \sigma I$.

(2) Rescale the observed data matrix: in approach (1), restrict the transformation matrix A to being a diagonal scaling matrix. If the factor model holds, there always exists such a scaling.

(3) Adding a matrix to correlation matrix: find matrix B such that $R + B = X + \sigma I$ with some constraints on B matrix.

Conclusion

Dohono and Gavish's method works well when its model assumptions as well as large p and N are met. It breaks down otherwise, although we found a surprising robustness to the method to violation of the equal residual variance assumption when p and N are large. Generally speaking, the eigenvalue-greater-than-one criterion is still an effective hard threshold to determine how many factors to retain, although DG's method can be better when N is very small and p is very large. DG's method is asymptotically good for recovering a latent covariance matrix or selecting the number of factors in factor analysis with increasing matrix size relative to number of factors.

References

[1] Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. Psychological Bulletin, 103(2), 276-279.

[2]Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151.

[3] Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19, 149–161.

[4] Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99(3), 432-442.

[5] Donoho, D. L., & Gavish, M. (2013). The optimal hard threshold for singular values is 4 / $\sqrt{3}$. Stanford University Statistics Department Technical Report 2013-04. arXiv: 1305.5870.

[6] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrica, 30(2), 179-185.

[7] Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

[8] Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41, 321–327.

[9] Nunnally, J. C., & Bernstein, I.H. (1994). Psychometric theory (3rd ed). New York: McGraw-Hill.