Contingency Table Analysis via Matrix Factorization

Kumer Pial Das [1], Jay Powell [2], Myron Katzoff [3], S. Stanley Young [4]

[1]Department of Mathematics,Lamar University, TX

[2]Better Schooling Systems, Pittsburgh, PA

[3]Center for Disease Control-Retired

[4]National Institute of Statistical Sciences

May 10, 2013

Outline

One of the goals of the OCER working group

"Singular value decompositions can be used to simultaneously cluster the rows (patients) and columns (variables) of a two-way table. New research indicates that non-negative matrix factorization, NMF, has distinct advantages in the identification of subgroups with distinct etiologies, Lee and Seung (1999). Scores computed via NMF might be better clustering variables for Local Control analysis."

Introduction

- Contingency tables of numeric data are often analyzed using dimension reduction methods like the singular value decomposition (SVD), and principal component analysis (PCA).
- This analysis produces score and loading matrices representing the rows and the columns of the original table and these matrices may be used for both prediction purposes and to gain structural understanding of the data.

Introduction

- Contingency tables of numeric data are often analyzed using dimension reduction methods like the singular value decomposition (SVD), and principal component analysis (PCA).
- This analysis produces score and loading matrices representing the rows and the columns of the original table and these matrices may be used for both prediction purposes and to gain structural understanding of the data.

Extracting factors from a contingency table

- IJ Good (1969) proposed using SVD to extract factors from a contingency table.
- We think Non-negative Matrix Factorization (NMF) will likely produce better results and interpretation.

Extracting factors from a contingency table

- IJ Good (1969) proposed using SVD to extract factors from a contingency table.
- We think Non-negative Matrix Factorization (NMF) will likely produce better results and interpretation.

Non-negative Matrix Factorization

- In some tables, the data entries are necessarily non-negative, and so the matrix factors meant to represent them should arguably also contain only non-negative elements.
- We describe here the use of NMF, an algorithm based on decomposition by parts that can reduce the dimension of data.

Non-negative Matrix Factorization

- In some tables, the data entries are necessarily non-negative, and so the matrix factors meant to represent them should arguably also contain only non-negative elements.
- We describe here the use of NMF, an algorithm based on decomposition by parts that can reduce the dimension of data.

Non-negative Matrix Factorization

- Let A be a $n \times p$ non-negative matrix, and $k > 0$ an integer. NMF consists in finding an approximation,

$$A \approx WH$$

  where W, H are $n \times k$ and $k \times p$ non-negative matrices, respectively.
- Columns of W are the underlying basis vectors.
- Columns of H give the weights associated with each basis vector.

Non-negative Matrix Factorization

- Let A be a $n \times p$ non-negative matrix, and $k > 0$ an integer. NMF consists in finding an approximation,

$$A \approx WH$$

  where W, H are $n \times k$ and $k \times p$ non-negative matrices, respectively.
- Columns of W are the underlying basis vectors.
- Columns of H give the weights associated with each basis vector.

Non-negative Matrix Factorization

- Let A be a $n \times p$ non-negative matrix, and $k > 0$ an integer. NMF consists in finding an approximation,

$$A \approx WH$$

  where W, H are $n \times k$ and $k \times p$ non-negative matrices, respectively.
- Columns of W are the underlying basis vectors.
- Columns of H give the weights associated with each basis vector.

- In practice, the factorization rank k is often chosen such that $k << min(n, p)$. In general, k can be bounded as $(n + p)k < np$. The objective behind this choice is to summarize and split the information contained in A into k factors: the columns of W.

- Depending on the application field, these factors are given different names:basis images, metagenes, source signals. We'll be using the term *source signals* in this presentation.

- In practice, the factorization rank k is often chosen such that $k << min(n, p)$. In general, k can be bounded as $(n + p)k < np$. The objective behind this choice is to summarize and split the information contained in A into k factors: the columns of W.

- Depending on the application field, these factors are given different names:basis images, metagenes, source signals. We'll be using the term *source signals* in this presentation.

Objective

- We study the use of PCA, SVD, and NMF to reduce the dimensionality of count data presented in a contingency table. Our primary goal is to remove noise and uncertainty. In theory, NMF can also be used for better interpretation of factoring matrices.

- As the rank k increases the method uncovers substructures, whose robustness can be evaluated by a cophenetic correlation coefficient. These substructures may also give evidence of nesting subtypes. Thus, NMF can reveal hierarchical structure when it exists but does not force such structure on the data.

Objective

- We study the use of PCA, SVD, and NMF to reduce the dimensionality of count data presented in a contingency table. Our primary goal is to remove noise and uncertainty. In theory, NMF can also be used for better interpretation of factoring matrices.

- As the rank k increases the method uncovers substructures, whose robustness can be evaluated by a cophenetic correlation coefficient. These substructures may also give evidence of nesting subtypes. Thus, NMF can reveal hierarchical structure when it exists but does not force such structure on the data.

Cophenetic Correlation Coefficient

- The cophenetic correlation coefficient is based on the consensus matrix (i.e. the average of connectivity matrices) and was proposed by Brunet et al. (2004) to measure the stability of the clusters obtained from NMF.

- It is defined as the Pearson correlation between the samples distances induced by the consensus matrix, seen as a similarity matrix and their cophenetic distances from a hierarchical clustering based on these very distances, by default an average linkage is used.

Cophenetic Correlation Coefficient

- The cophenetic correlation coefficient is based on the consensus matrix (i.e. the average of connectivity matrices) and was proposed by Brunet et al. (2004) to measure the stability of the clusters obtained from NMF.

- It is defined as the Pearson correlation between the samples distances induced by the consensus matrix, seen as a similarity matrix and their cophenetic distances from a hierarchical clustering based on these very distances, by default an average linkage is used.

A Typical Genomic Study

- We may consider a data set consisting of the expression levels of M genes in N samples. For gene expression studies, M is typically in the thousands, and N is typically <100
- The data can be represented by an expression matrix A of size $M \times N$.
- One possible goal could be to find a small number of metagenes, each defined as a positive linear combination of the M genes. We can then approximate the gene expression pattern of samples as positive linear combinations of these metagenes.

A Typical Genomic Study

- We may consider a data set consisting of the expression levels of M genes in N samples. For gene expression studies, M is typically in the thousands, and N is typically <100
- The data can be represented by an expression matrix A of size $M \times N$.
- One possible goal could be to find a small number of metagenes, each defined as a positive linear combination of the M genes. We can then approximate the gene expression pattern of samples as positive linear combinations of these metagenes.

A Typical Genomic Study

- We may consider a data set consisting of the expression levels of M genes in N samples. For gene expression studies, M is typically in the thousands, and N is typically <100
- The data can be represented by an expression matrix A of size $M \times N$.
- One possible goal could be to find a small number of metagenes, each defined as a positive linear combination of the M genes. We can then approximate the gene expression pattern of samples as positive linear combinations of these metagenes.

Alzheimer Data Set

- National Institute on Aging (NIA) and NIH are conducting research by collecting data from 29 NIA funded Alzheimer's Disease Centers.
- There are approximately 25,000 subjects.
- The data set includes more than 700 variables, representing demographics, behavioral status, cognitive testing, medical history, family history, clinical impressions, and diagnoses.

Alzheimer Data Set

- National Institute on Aging (NIA) and NIH are conducting research by collecting data from 29 NIA funded Alzheimer's Disease Centers.
- There are approximately 25,000 subjects.
- The data set includes more than 700 variables, representing demographics, behavioral status, cognitive testing, medical history, family history, clinical impressions, and diagnoses.

Alzheimer Data Set

- National Institute on Aging (NIA) and NIH are conducting research by collecting data from 29 NIA funded Alzheimer's Disease Centers.
- There are approximately 25,000 subjects.
- The data set includes more than 700 variables, representing demographics, behavioral status, cognitive testing, medical history, family history, clinical impressions, and diagnoses.

2013 Geoffrey Beene Global NeuroDiscovery Challenge

- The Foundation for the National Institute of Health (FNIH), in association with the Geoffrey Beene Foundation Alzheimer's initiative, is conducting a study to identify male/female differences in the early cognitive decline that precedes Alzheimer's disease.
- The 2013 Geoffrey Beene Global NeuroDiscovery Challenge calls for original and innovative hypotheses on gender differences in pathology and neuro degenerative decline leading to Alzheimer's disease.

2013 Geoffrey Beene Global NeuroDiscovery Challenge

- The Foundation for the National Institute of Health (FNIH), in association with the Geoffrey Beene Foundation Alzheimer's initiative, is conducting a study to identify male/female differences in the early cognitive decline that precedes Alzheimer's disease.
- The 2013 Geoffrey Beene Global NeuroDiscovery Challenge calls for original and innovative hypotheses on gender differences in pathology and neuro degenerative decline leading to Alzheimer's disease.

2013 Geoffrey Beene Global NeuroDiscovery Challenge

- The winning submission(s) will receive $100,000 in prize awards.
- For more information you may visit www.geoffreybeenechallenge.org.

2013 Geoffrey Beene Global NeuroDiscovery Challenge

- The winning submission(s) will receive $100,000 in prize awards.
- For more information you may visit www.geoffreybeenechallenge.org.

Description of the data set

- J. Powell and his colleagues designed various education studies. Most of these studies originated with an observation made in the early 1960's wherein some "wrong" answers given to multiple tests appeared to reflect systematic selection by test subjects.

- One of these studies has used a multiple choice test (known as "the Proverbs Test") contains 40 items, each with 4 alternatives. The test was given twice in each year. We have added another alternative for missing data.

Description of the data set

- J. Powell and his colleagues designed various education studies. Most of these studies originated with an observation made in the early 1960's wherein some "wrong" answers given to multiple tests appeared to reflect systematic selection by test subjects.

- One of these studies has used a multiple choice test (known as "the Proverbs Test") contains 40 items, each with 4 alternatives. The test was given twice in each year. We have added another alternative for missing data.

Netherlandish Proverbs by P. Bruegel the Elder, 1559

Q32: Don't cast pearls before swine

1. Put your efforts where they are appreciated
2. Don't give pearl to fools
3. Don't be wasteful
4. Don't always put yourself before everybody

Age Group	1	2	3	4
1	39	42	39	51
2	33	37	50	54
3	56	30	43	44
4	57	24	38	56
5	50	15	56	56
6	66	10	35	62
7	74	12	40	48
8	89	4	24	56
9	99	3	27	46
10	85	9	29	51
11	89	4	28	52
12	101	2	22	50
13	121	1	23	31

All 40 questions

- The Windsor education data is represented by an expresssion matrix A of size 2293×160 whose rows contain 2,293 students and coulmn contains 160 variables (40 questions with four alternatives).
- Matrix W has size $2293 \times k$, with each of the k columns defining a *source signal*.
- Matrix H has size $k \times 160$, with each of the 160 columns representing a *source signal* expression pattern of the corresponding sample.

All 40 questions

- The Windsor education data is represented by an expresssion matrix A of size 2293×160 whose rows contain 2,293 students and coulmn contains 160 variables (40 questions with four alternatives).
- Matrix W has size $2293 \times k$, with each of the k columns defining a *source signal*.
- Matrix H has size $k \times 160$, with each of the 160 columns representing a *source signal* expression pattern of the corresponding sample.

All 40 questions

- The Windsor education data is represented by an expresssion matrix A of size 2293×160 whose rows contain 2,293 students and coulmn contains 160 variables (40 questions with four alternatives).
- Matrix W has size $2293 \times k$, with each of the k columns defining a *source signal*.
- Matrix H has size $k \times 160$, with each of the 160 columns representing a *source signal* expression pattern of the corresponding sample.

Estimating the Factorization Rank

- A critical parameter in NMF is the factorization rank k. It defines the number of *source signals* used to approximate the target matrix.
- Given a NMF method and the trgaet matrix, a common approach of deciding on k is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria.
- We use both R and JMP to estimate the value of k.

Estimating the Factorization Rank

- A critical parameter in NMF is the factorization rank k. It defines the number of *source signals* used to approximate the target matrix.
- Given a NMF method and the trgaet matrix, a common approach of deciding on k is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria.
- We use both R and JMP to estimate the value of k.

Estimating the Factorization Rank

- A critical parameter in NMF is the factorization rank k. It defines the number of *source signals* used to approximate the target matrix.
- Given a NMF method and the trgaet matrix, a common approach of deciding on k is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria.
- We use both R and JMP to estimate the value of k.

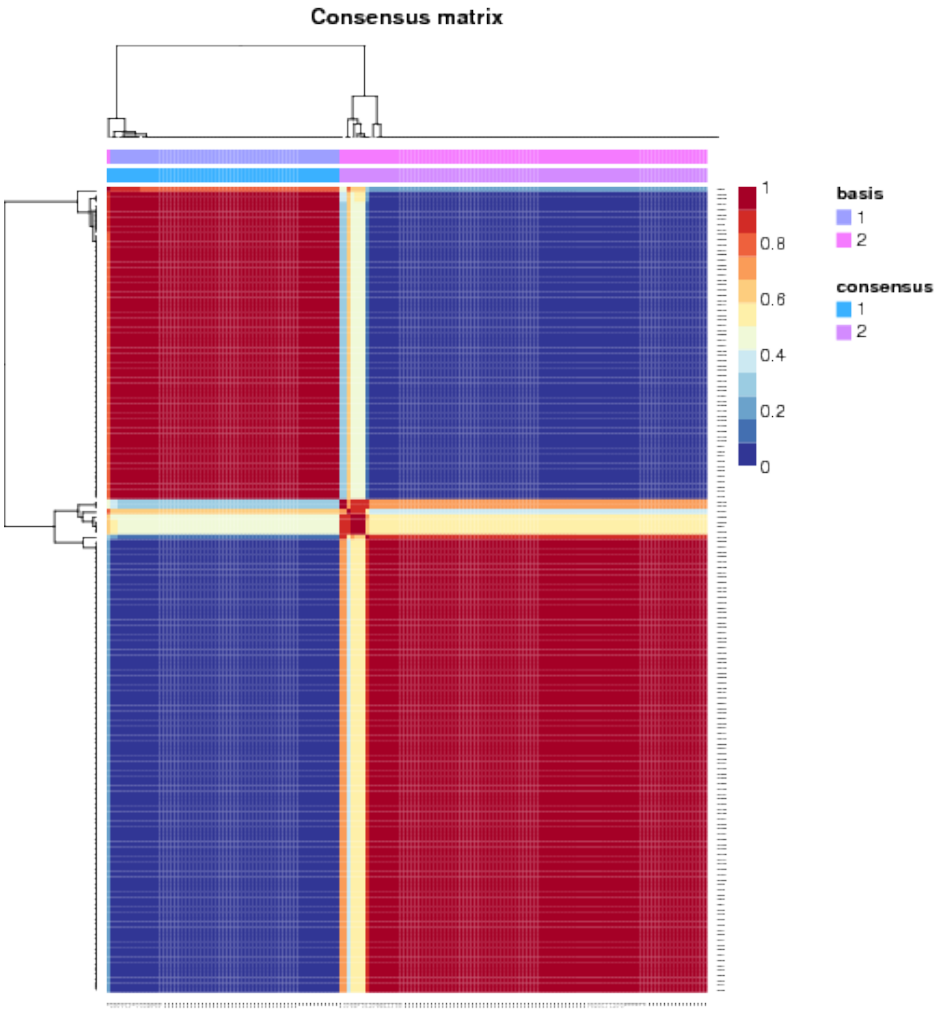
Interpreting the consensus matrix

- The consensus matrices are computed at $k = 2$ to 10 for the Windsor education data.
- Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0(deep blue, samples are never in the same cluster) to 1 (dark red, samples are always in the same cluster).

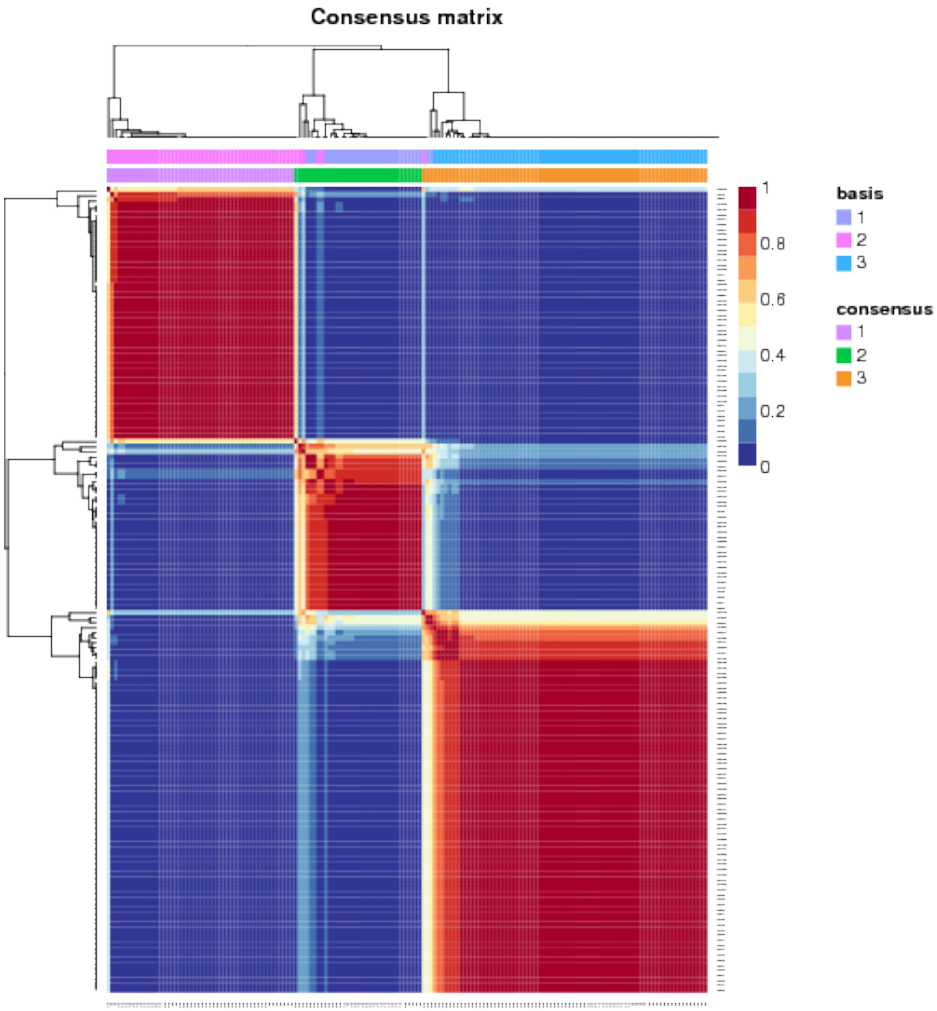
Interpreting the consensus matrix

- The consensus matrices are computed at $k = 2$ to 10 for the Windsor education data.
- Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0(deep blue, samples are never in the same cluster) to 1 (dark red, samples are always in the same cluster).

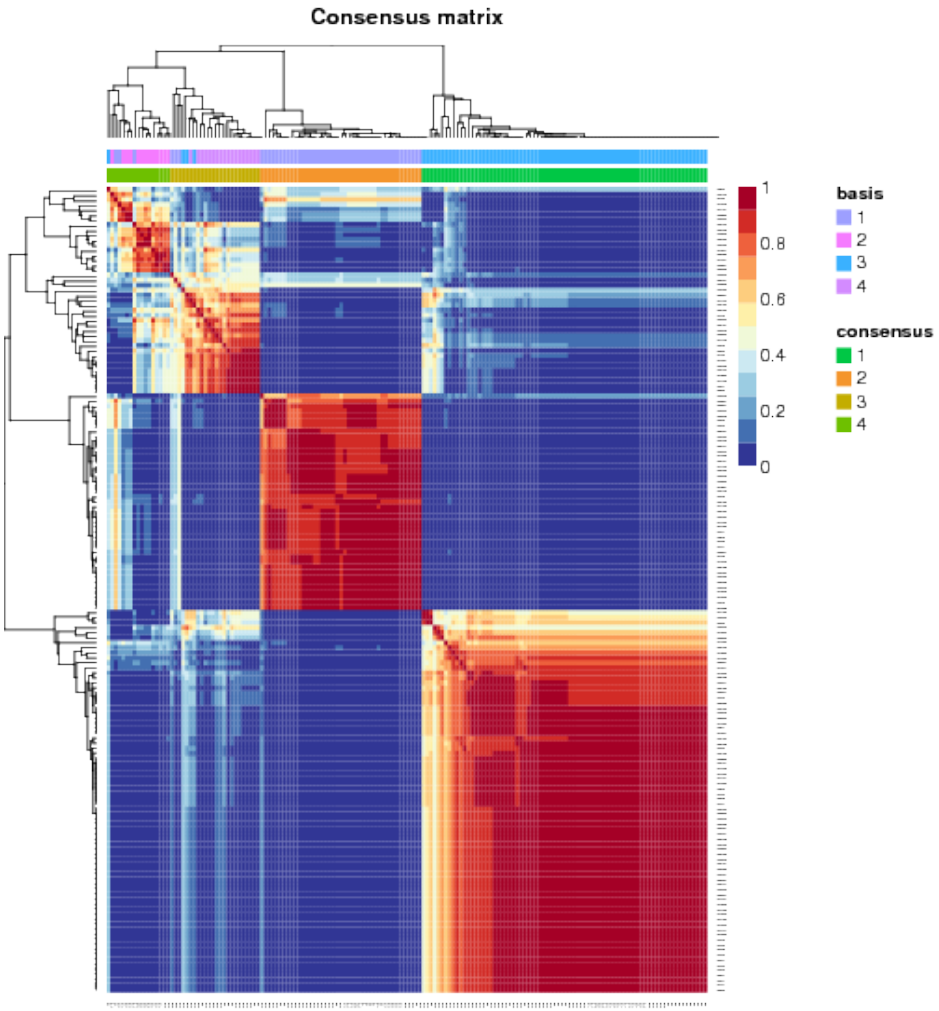
Consensus Map of all questions for rank 2



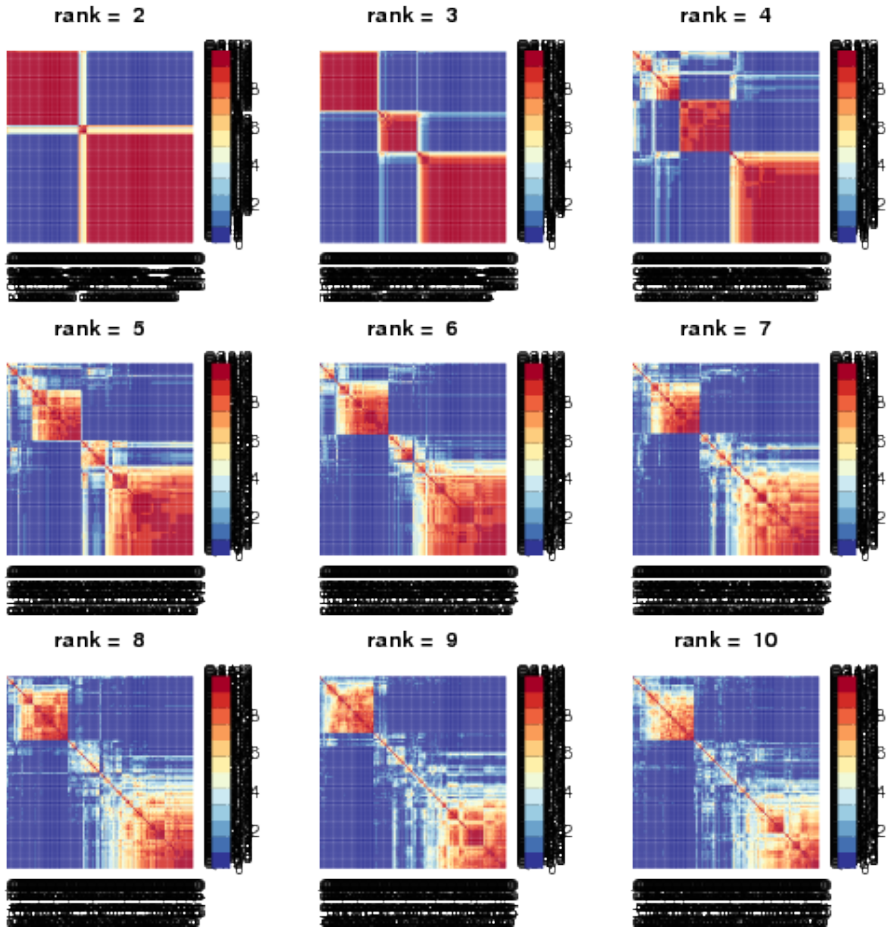
Consensus Map of all questions for rank 3



Consensus Map of all questions for rank 4



Consensus Map of all questions for rank 2-10



Cophenetic Correlation Coefficeint

- Although a visual inspection is important, it's also important to have quantitative measure of stability for each value of k. One measure proposed by (Brunet, 2004) is cophenetic coefficient, which indicates the dispersion of the consensus matrix, defined as the average connectivity matrix over many clustering runs.
- We observe how cophenetic coefficient changes as k increases. We select values of k where the magnitude of cophenetic coefficient begins to fall.
- (Hutchins et al. 2008) suggested to choose the first value where the RSS curve presents an inflection point.

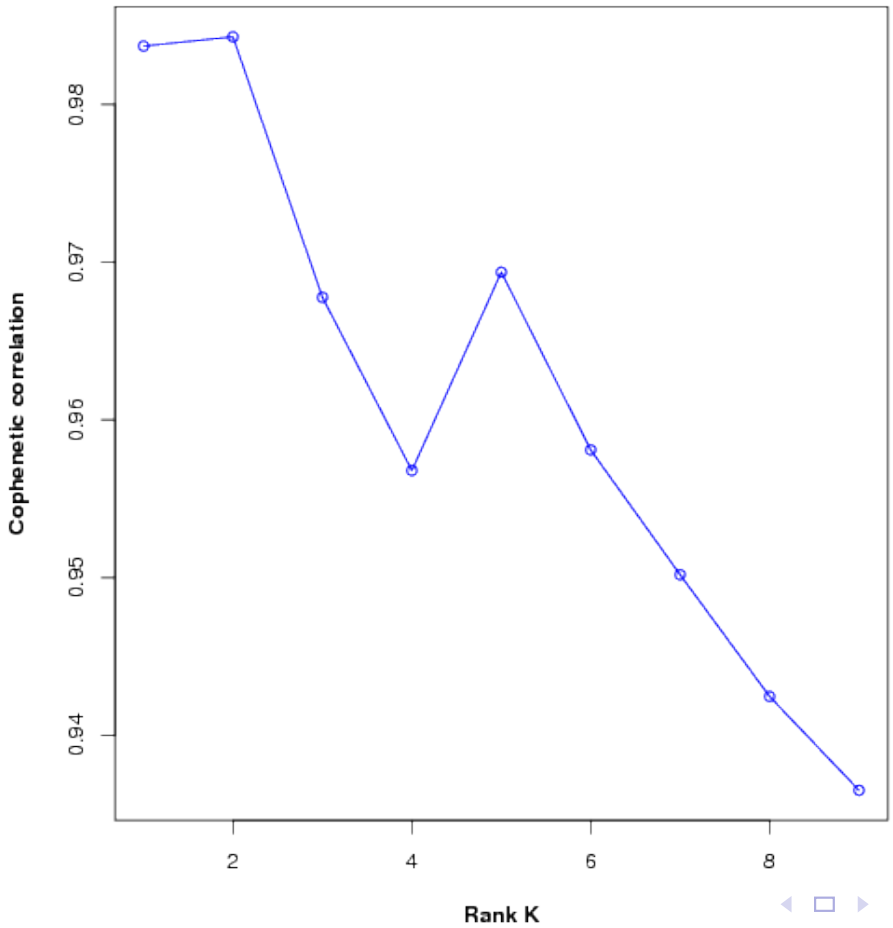
Cophenetic Correlation Coefficeint

- Although a visual inspection is important, it's also important to have quantitative measure of stability for each value of k. One measure proposed by (Brunet, 2004) is cophenetic coefficient, which indicates the dispersion of the consensus matrix, defined as the average connectivity matrix over many clustering runs.
- We observe how cophenetic coefficient changes as k increases. We select values of k where the magnitude of cophenetic coefficient begins to fall.
- (Hutchins et al. 2008) suggested to choose the first value where the RSS curve presents an inflection point.

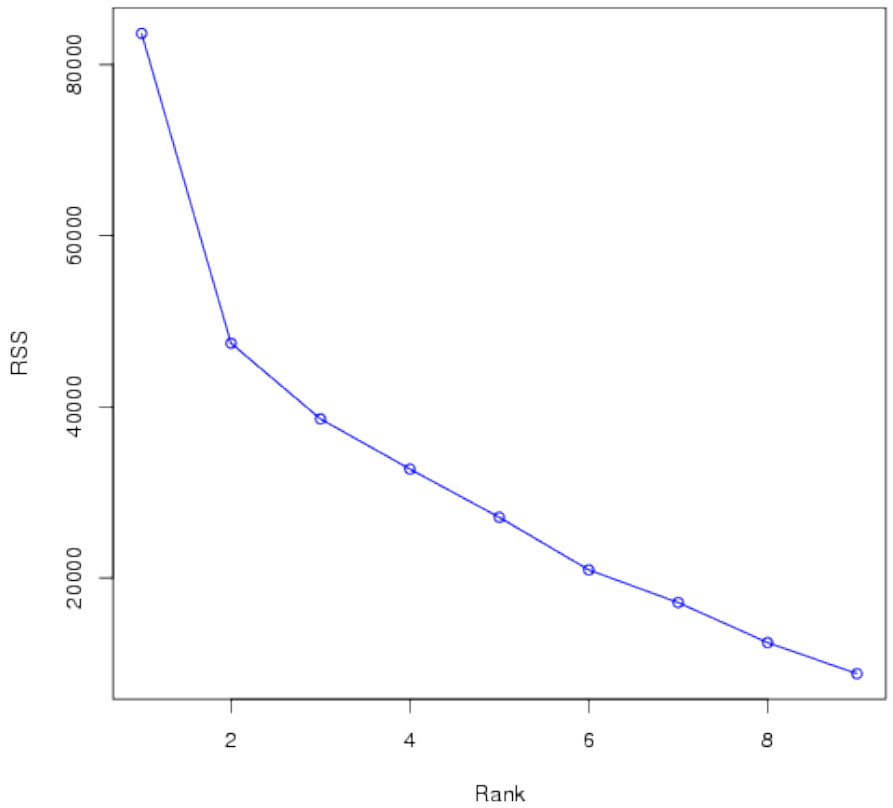
Cophenetic Correlation Coefficeint

- Although a visual inspection is important, it's also important to have quantitative measure of stability for each value of k. One measure proposed by (Brunet, 2004) is cophenetic coefficient, which indicates the dispersion of the consensus matrix, defined as the average connectivity matrix over many clustering runs.
- We observe how cophenetic coefficient changes as k increases. We select values of k where the magnitude of cophenetic coefficient begins to fall.
- (Hutchins et al. 2008) suggested to choose the first value where the RSS curve presents an inflection point.

Estimating *k*



Estimating *k*



Concluding Remarks

- NMF is looking at all answer columns without paying attention to the natural grouping of 4 answers for each question.
- NMF does not "know" the correct answer. NMF is "unsupervised".
- NMF is looking at how students of different ages answer proverb questions. It is looking at the shape of the age response curve.

Concluding Remarks

- NMF is looking at all answer columns without paying attention to the natural grouping of 4 answers for each question.
- NMF does not "know" the correct answer. NMF is "unsupervised".
- NMF is looking at how students of different ages answer proverb questions. It is looking at the shape of the age response curve.

χ^2 partitioning

- We group this data by the usual χ^2 partitioning approach. But we have found that the problem of dimension reduction remains.
- We believe this is where NMF should be useful.
- Ideally, the factors that distinguish the members of the ultimate groups are the same but why must that be the case? Some factors may be active in one group that are not actors in another; but what are they? We think NMF answers that question.

χ^2 partitioning

- We group this data by the usual χ^2 partitioning approach. But we have found that the problem of dimension reduction remains.
- We believe this is where NMF should be useful.
- Ideally, the factors that distinguish the members of the ultimate groups are the same but why must that be the case? Some factors may be active in one group that are not actors in another; but what are they? We think NMF answers that question.

χ^2 partitioning

- We group this data by the usual χ^2 partitioning approach. But we have found that the problem of dimension reduction remains.
- We believe this is where NMF should be useful.
- Ideally, the factors that distinguish the members of the ultimate groups are the same but why must that be the case? Some factors may be active in one group that are not actors in another; but what are they? We think NMF answers that question.

Concluding Remarks

- What's after PCA/SVD/NMF?
- A possible way forward is to use NMF to get response variables for recursive partitioning or local control approach.

Concluding Remarks

- What's after PCA/SVD/NMF?
- A possible way forward is to use NMF to get response variables for recursive partitioning or local control approach.

References

- Brunet, J.P., Tamayo, P., Gould, T.R., and Mesirov, J.P.(2004) Metagenes and Molecular Pattern Discovery using Matrix Factorization, Proceedings of the National Academy of Sciences. 101,4164-4169.
- Good, I.J.(1969) Some Applications of the Singular Decomposition of a Matrix, Technometrics. 11(4),823-831.
- Lee, D.D., and Seung, H.S.(1999) Algorithms for Non-negative Matrix Factorization, Nature. 401,788-793.

Thank You

Thank You

Questions