Non-negative Matrix Factorisation (Lee, Seung 1999)

Mohammad Sohaib Ahmad (MSA53) and Junjie Pan (JP697)

Overview of Paper

"Learning the parts of objects by non-negative matrix factorization" applies the method of matrix factorisation in a non- negative setting. The paper finds that the added non-negative constraint in a dimensionality reduction type of problem results in a representation by parts form of learning, as opposed to a linear combination of the data set (from e.g. VQ or PCA).

This paper was seminal in igniting research into this area and in particular, they highlighted the possible links to human memory and perception, which as they say, could be thought as parts-based representational learning: "There is psychological and physiological evidence for parts-based representations in the brain, and certain computational theories of object recognition rely on such representations."

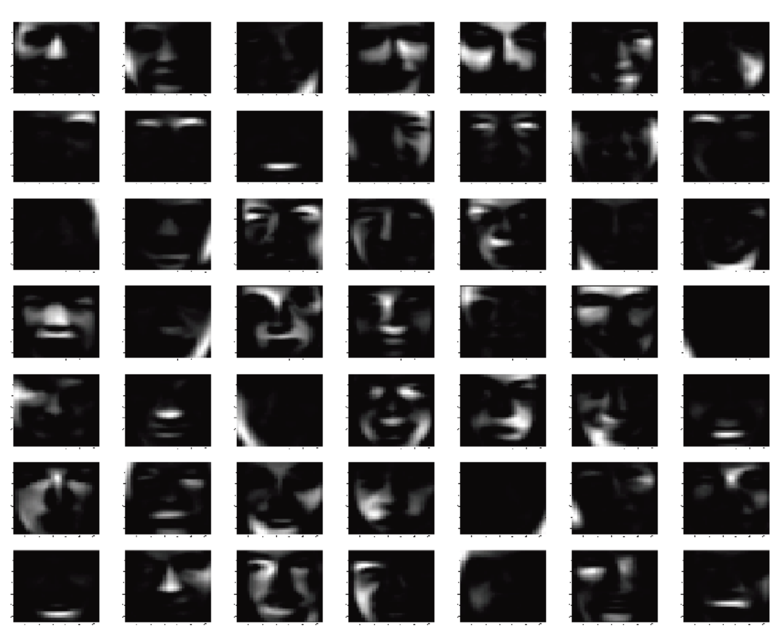
Note: Formulae's for both methods are provided in Formulae's corner. Please go over these and ensure that these are understood before progressing.

Replicating Results (1)

NMF Decomposition of Images

The figure on the right are the grid results from NMF.

- NMF results have representational feel
- Most prominent features are highlighted first:
  - Eyes
  - Nose
  - The T-zone area
  - Cheeks.
- Encoded images are sparse, large black regions.
  - These basis images can have a sparse representation of features because they are non-global features.
  - Different types of each feature can also be captured by NMF. This is because the different columns of the weight matrix capture different features and given the non-negativity constraint, images can only be reconstructed through additive processes, which is how NMF learns a parts-based representation.

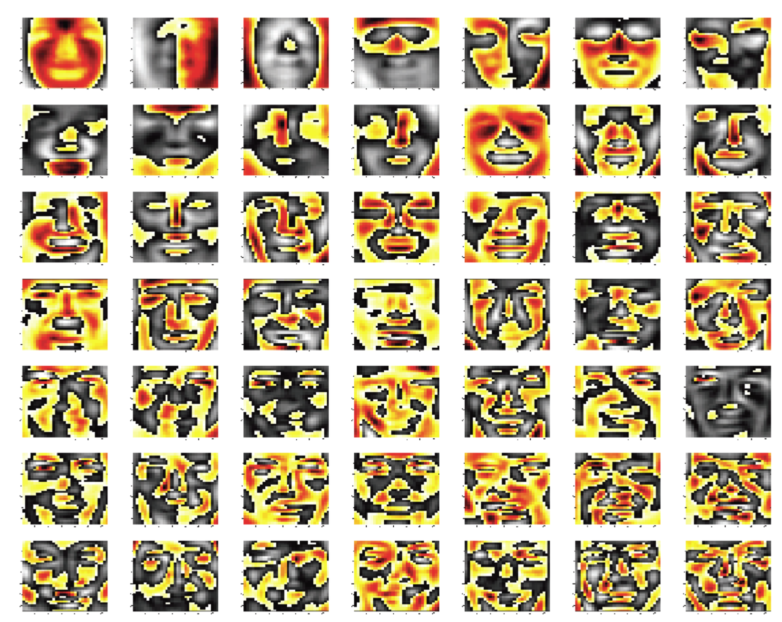


Replicating Results (2)

PCA Decomposition of Images

The figure on the right are the grid results from NMF.

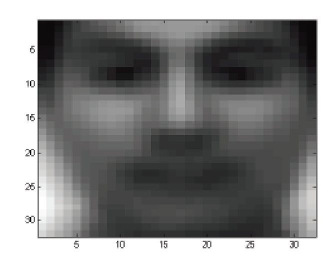
- Same Initial "V" Matrix as NMF
- PCA constrains the matrix to be orthogonal
- Decomposed Images are a linear combination of images in the dataset,.
- Encoded images have no non-negativity constraint, but results of the orthogonal-decomposed matrix makes little sense:
  - The blurry type of images that are created are often referenced as 'eigenfaces'
  - 'Eigenfaces' are the decomposed images in the direction of largest variance. Few of the images have an obvious visual interpretation - if anything, none do.



Deriving an "Average Looking Person" from the dataset

As a quick extension, we thought it might be interesting to derive the average looking face from the dataset that we were using, by using the following constraint:

- W = N x r, where r is constrained to be 1 feature.



The above is a visual implementation where starting from the initial matrix V that contains all the information regarding the pixels and images, these are then decomposed into the W and H matrices, where we limit the number of features to 1 (r=1).

NMF And Topic Modelling

- The figure on the right is an implementation of Topic Modelling on the a given set of semantic features from a dataset taken from the Encyclopaedia Britannica.
- We ran with 450 semantic features (columns of W).
- As these vectors are high dimensional (of length of 1,444), we picked the weights for 4 features with highest weights.
- In this table, each of the semantic features is represented by a list of ten words with the highest frequency. The numbers in this table correspond to the weights for the corresponding semantic features.
- Note that this final word count vector was approximated by a superposition that gave high weight to the upper two semantic features, and lower weight to the lower two.

Feature	Weights	Feature	Weights
mines	47.825%	circuit	42.585%
electrical	9.884%	line	15.753%
used	9.788%	connected	13.060%
defence	8.339%	battery	9.716%
shore	8.320%	connexion	8.572%
class	7.430%	telephone	8.268%
field	6.718%	placed	6.713%
charge	6.699%	position	6.108%
station	6.591%	arrangement	5.375%
apparatus	6.546%	associated	5.274%
water	79.201%	surface	57.805%
taken	3.026%	curve	11.459%
generally	2.572%	normal	8.696%
rise	1.917%	form	8.445%
long	1.799%	systems	6.611%
nearly	1.162%	having	6.123%
surface	0.937%	lines	5.692%
tidal	0.911%	theory	5.129%
containing	0.892%	particular	5.119%
bar	0.842%	point	5.086%

Feature	Counts
Mines	57
Firing	20
Used	12
Electrical	12
Defence	10
Shore	10
Class	9
Charge	8
Small	8
Apparatus	8

Formulae Corner

PCA

- PCA is the orthogonal linear transformation of a coordinate system such that the data with the greatest variance comes to lie on the first coordinate by some projection.
- The basic process to run principal component analysis has been written on the right hand side, where the covariance matrix is first decomposed, following by then deriving the eigenvalues of the covariance matrix.
- From here, existing data is amended to make it orthogonal and to highlight key features.

Given a data set of N centered observations in a d-dimensional space

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}, \Sigma_{k=1}^{N}\mathbf{x}_k = 0, \mathbf{x}_k \in R^d$$

- PCA diagonalizes the covariance matrix:

$$C = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k\mathbf{x}_k^T$$

- It is necessary to solve the following system of equations:

$$\lambda\mathbf{v} = C\mathbf{v} \Leftrightarrow$$

$$\lambda(\mathbf{x}_k \cdot \mathbf{v}) = (\mathbf{x}_k \cdot C\mathbf{v}), \forall k = 1, 2, \ldots, N.$$

- We can define the same computation in another dot product space F:

$$\phi : R^d \to F, \mathbf{x} \mapsto \mathring{\mathbf{x}}$$

NMF

- NMF is the process by which we decompose a matrix into two other matrices, following the example of the first image and the first equation.
- NMF does not have a closed form solution, so requires the maximisation of the given objective function.
- There is also a constraint on the number of features that we can attempt to model, such that the following relation must hold: (N+M)*r<N*M, where r is the number of features.
- Using the final update equations, we employ the method of steepest descent to converge to the local maximum.
- Note: the exact form of the object function is not as crucial as the non-negativity constraint for the success of this method

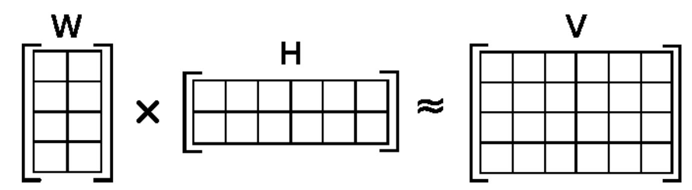


Illustration of approximate non-negative matrix factorization: the matrix V is represented by the two smaller matrices W and H, which, when multiplied, approximately reconstruct V.

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia}H_{a\mu}$$

$$F = \sum_{i=1}^{n}\sum_{\mu=1}^{m}[V_{i\mu}\log(WH)_{i\mu} - (WH)_{i\mu}]$$

$$\boxed{\begin{gathered} W_{ia} \leftarrow W_{ia}\sum_{\mu}\frac{V_{i\mu}}{(WH)_{i\mu}}H_{a\mu} \\ W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}} \end{gathered}}$$

$$\boxed{H_{a\mu} \leftarrow H_{a\mu}\sum_i W_{ia}\frac{V_{i\mu}}{(WH)_{i\mu}}}$$

References

1. DD Lee, HS Seung. 1999 "Learning the parts of objects by non-negative matrix factorization"
2. Dempster, A. P., Laired, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm.
3. Encyclopaedia Britannica Eleventh Edition
4. Reference for list of common ("Stop") words, referenced from: http://xpo6.com/list-of-english-stop-words/