

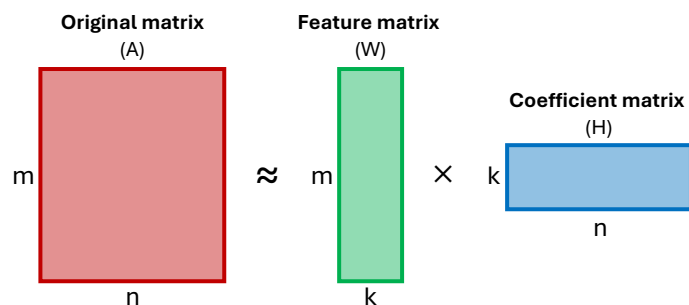
nmf: AN IMPLEMENTATION OF NON-NEGATIVE MATRIX FACTORISATION IN STATA

JA Batty (Wellcome Trust Clinical PhD Fellow)^{1,2} & M Hall (Associate Professor of Epidemiology)^{1,2}

Affiliations: ¹Leeds Institute for Data Analytics, University of Leeds, Leeds, UK, ²Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK

BACKGROUND AND MOTIVATION: WHAT IS NMF?

- An **unsupervised machine learning approach**: used for **dimensionality reduction, clustering** and **feature extraction**.
- Decomposes** matrix **A** into **lower-rank matrices W** and **H**, such that $A \approx WH$.
- A**, **W** and **H** are **non-negative, real numbers**. The desired rank ($k < \min(m, n)$)



- NMF has been applied in several contexts: **text mining, recommendation systems, computer vision, signal processing** and **bioinformatics**.
- It has **not** entered **widespread use** in **biostatistics** or **epidemiology**.

IMPLEMENTATION: HOW DOES NMF WORK?

- NMF is an **NP-hard** problem; in general, **no exact solution exists**.
- Matrices **W** and **H** are initialised with **random** or **calculated** (SVD) values, ≥ 0 .
- These values are updated using **multiplicative update rules** first reported by Paatero and Tapper[1] and popularised by Lee and Seung[2, 3].

$$H \leftarrow H \circ \frac{W^T X}{W^T H} \quad W \leftarrow W \circ \frac{X}{W H^T}$$

- Optimisation** is performed by **minimising** $\|A - WH\|$. **3 options for loss functions**: (1) Frobenius (Euclidean), (2) Kullback-Leibler, and (3) Itakura-Saito[4].
- Multiple iterations** are performed until **convergence** (or a fixed no. epochs).
- Missing values** are permitted in **A**.
- NMF algorithm implemented using **Mata** for **speed** and **stability**.

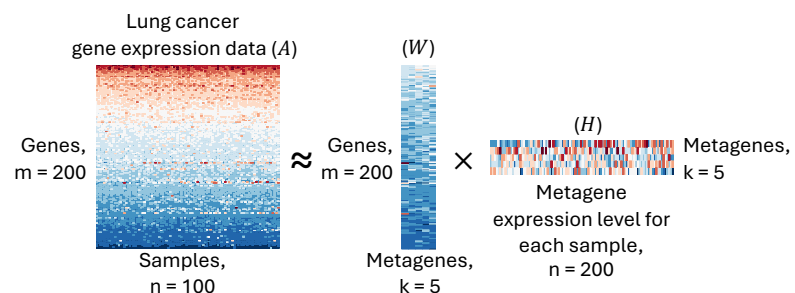
[1] Paatero, P & Tapper U, Environmetrics (1994). [2] Lee, D. & Seung, H, Nature (1999).

[3] Lee, D. & Seung, H, NeurIPS (2000).

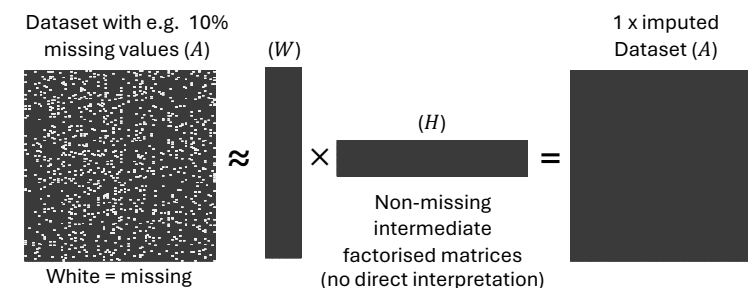
[4] Févotte & Idier, Neurat Comput. (2011).

EXAMPLES USING NMF IN STATA

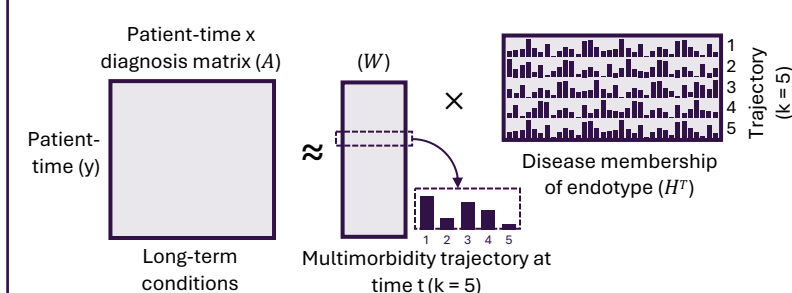
- Identification of metagenes from gene expression data** (metagenes.do; nsclc.dta) → reduction of 200 genes to 5 ‘metagenes’



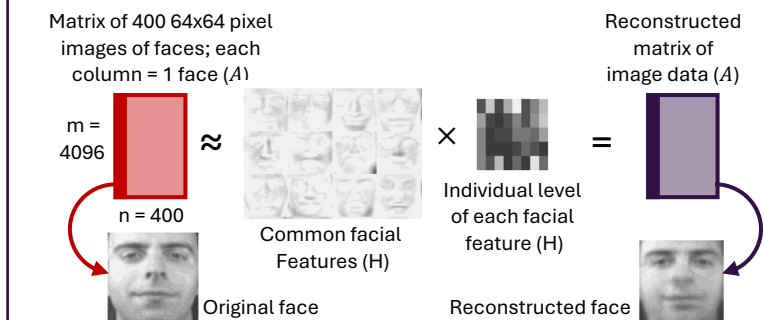
- Imputation of missing data** (imputation.do; nsclc.dta) → multiplication results in ‘complete’ **W** and **H**



- Identification of disease trajectories** (trajectories.do; disease_matrix.dta) → performing temporal clustering



- Identification of facial features from photos of faces** (imputation.do; faces.dta) → identifying lower-level features of images



USING NMF IN STATA: QUICK START GUIDE

- Installation:** `ssc install nmf`
- Detailed help file:** `help nmf`
- Example syntax:** `nmf cols*, k(15) initial(randomu) epoch(100) /// stop(1.0e-4) loss(kl)`
- All examples / data:** <https://github.com/jonathanbatty/stata-nmf>

ACKNOWLEDGEMENTS

- This project was funded by:
- The AI for Science and Government Fund (via The Alan Turing Institute; TU/ASG/R-SPEH-114)
 - The Wellcome Trust (JB: 227498/Z/23/Z, MH: 206470/Z/17/Z).

The Alan Turing Institute

