Optimal recovery of missing values for non-negative matrix factorization: A probabilistic error bound

R. Chen [1] *L. R. Varshney* [1 2]

ICML Artemiss 2020

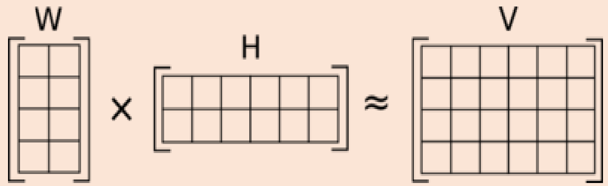
Goal

Quantify error of downstream processing after imputation (instead of imputation error itself)

Background

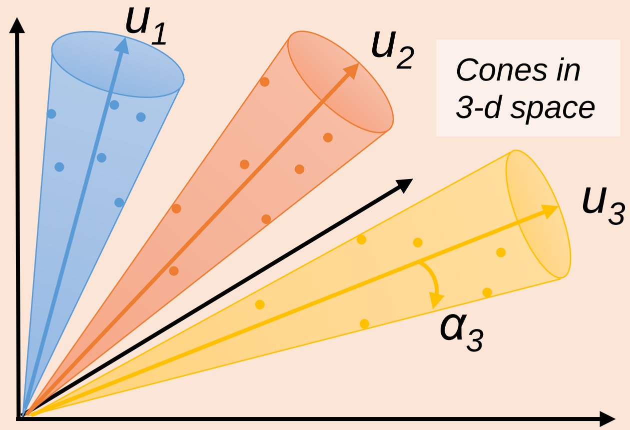
Non-negative matrix factorization (NMF)

- Scientists like NMF
- *N* samples, *F* observations each
- Given non-negative *F*×*N* matrix **V,** find non-negative factor matrices **W*** (*F*×*K*) and **H*** (*K*×*N*)



- **W*** contains *K* cluster prototypes $u_1,\ldots,u_k$

- If data is well-separated, we can represent a rank-1 NMF of our data as *K* well-separated[3] cones located in the non-negative orthant of an *F*-dimensional space



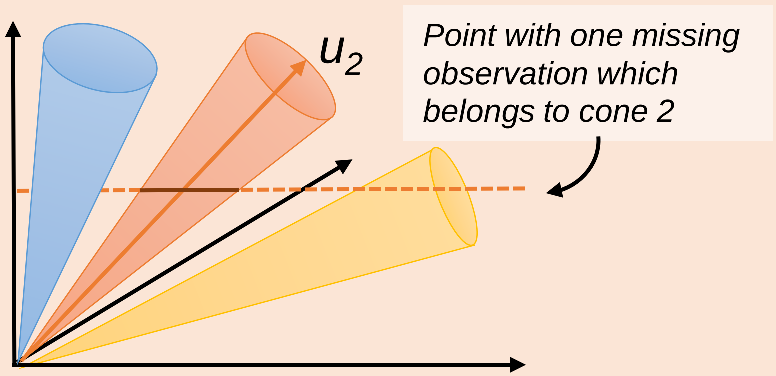
- The *N* data points are just noisy versions of the prototypes $u_1,\ldots,u_k$
- Size of cones given by $\alpha_1,\ldots,\alpha_k$
- Reconstruction error of a rank-1 NMF is given by:

$$\frac{\|\boldsymbol{V}-\boldsymbol{W}^*\boldsymbol{H}^*\|_F}{\|\boldsymbol{V}\|_F} \leq \max_k \sin \alpha_k$$

Imputation with Optimal Recovery

- Find *K* cones using data no missing values (fully-observed data)

- Impute using optimal recovery:
  - For *v* with missing values, determine feasible source cones. If more than one possible source cone, just pick one
  - Impute missing values with minimax center



[3] Well-separated means it's easy to find clusters

[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign [2] Salesforce Research, Palo Alto, CA, USA

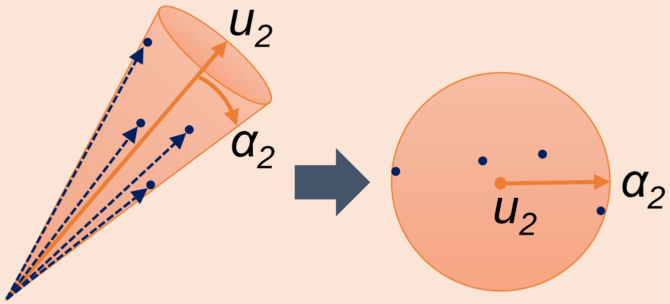
Optimal recovery of missing values for non-negative matrix factorization: A probabilistic error bound

R. Chen and L. R. Varshney

ICML Artemiss 2020

Setup

We can disregard the lengths of the data vectors and keep only the angle between the vector and the prototype. Thus, an *F*-dimensional cone can be viewed as an *(F-1)*-dimensional ball with radius α_k.



We calculate some *expected reconstruction errors* in Theorems 2 and 3 based on how the points are distributed in the ball, but in Theorem 1, we give an *upper bound*.

Theorem 1

Suppose we have *N* points drawn uniformly at random from *K* well-separated[1] cones, and points are distributed along the radius uniformly at random. Assume data is MCAR, but there is at least one fully-observed point per cluster. If we impute using optimal recovery, and we perform a rank-1 NMF to obtain ***W**** and ***H****, then the reconstruction error is

$$\frac{\|V - W^* H^*\|_F}{\|V\|_F} \leq \max_k \sin \alpha_k.$$

The error bound is the same as the result without missing values. This is because we are just calculating the error of the prototypes, or "cluster centers."

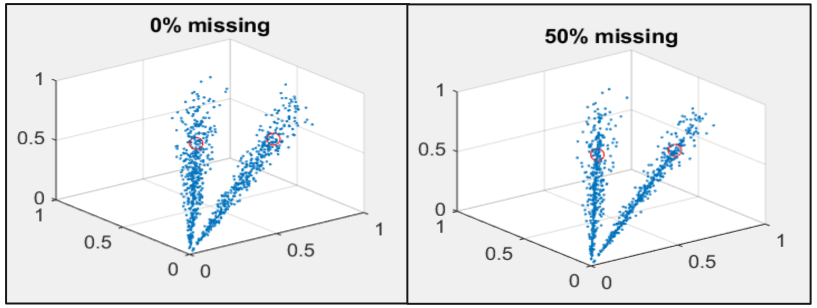


Illustration of optimal recovery imputation

Minimax and Fairness

Philosopher John Rawls argues that inequalities should only exist if they result in the worst off being better off. In a scenario where one's place in society is chosen at random (including social status and other assets), one would prefer to land in a society that plays by a minimax rule, where the disadvantage of the worst off is minimized.

Future Work

How does minimax imputation impact fairness in decision-making and clustering?