Error computation in matrix factorization problems

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : December 23, 2018
Last update : May 16, 2019

Overview

1. The factroization problem
   - Measuring the performance of a factorization
   - The non-uniqueness of factorization
2. Matching columns as an assignment problem
   - Estimation of the permutation matrix by correlation
   - Estimation of the permutation matrix by assignment problem
3. Summary

Factorization of a matrix

Consider we have a matrix factorization

$$X \approx WH$$

where

- $X \in \mathbb{R}^{m \times n}$ is the given matrix
- $W \in \mathbb{R}^{m \times r}$ and $H \in \mathbb{R}^{r \times n}$ are factors to approximate X
- m, n, r are known dimensions

The goal of the matrix factorization is to find two factor matrices W and H such that the product WH approximates the given X "well".

But how to define "well" ?
How to measure the "performance" of such factorization ?

The residue matrix

The performance of a factorization can be measured by the "distance" between the estimator WH towards X.

Definition. The **residue matrix** (or error matrix) E is the difference between X and the estimator WH :

$$E := X - WH$$

As E tells the difference between X and WH, hence the size of E tells the distance between the estimator WH and X

i.e. we say the factorization is "good" if the "size" of E is small.

Question : how to quantify the size of E ?

The size of error in F-norm

There are various way to quantify the "size" of E.

For example, using Frobenius norm, we have

$$\text{Performance} = \|E\|_F = \|X - WH\|_F$$

By the definition of F-norm $\|E\|_F = \sqrt{\sum_{ij} |E_{ij}|^2}$, what the above expression means :

taking all elements of E, which is the difference between X and WH into account with equal weight to generate a number that describe how big the error is.

In other words, we consider all elements of E have the same amount of **importance** in the calculation.

The size of error in other measures

It is possible that some elements in WH are more "important" than others. Hence weighted F-norm can be used

$$\|\Omega \bigodot E\|_F = \|\Omega \otimes (X - WH)\|$$

where Ω is a m-by-n matrix indicating the importance of each element E_{ij}, $\bigodot$ is element-wise product.

Other measures are :

- L_1 norm, for robustness
- L_2 norm (spectral norm)
- Nuclear norm, which is an approximation of rank function, it measure the "rank" of E
- KL divergence, beta-divergence . . . , which are used in audio research

Ground truth model

Assumes matrix X is generated in the following ground truth model

$$X = W_0 H_0 + N$$

where

- $W_0 \in \mathbb{R}^{m \times r}$ and $H_0 \in \mathbb{R}^{r \times n}$ are the true factors
- $N \in \mathbb{R}^{m \times n}$ is the additive noise matrix

If W_0 and H_0 are known, the performance of the factorization can also be measured as follows

$$\text{Performance} = \|\Delta_W\|_F = \|W_0 - W\|_F$$

$$\text{Performance} = \|\Delta_H\|_F = \|H_0 - H\|_F$$

Relationship between $\|E\|_F$ and $\|\Delta\|_F$

Δ_W and Δ_H are the error corresponding to W_0 and H_0 respectively :

$$W = W_0 + \Delta_W, \qquad H = H_0 + \Delta_W$$

Hence we have

$$WH = (W_0 + \Delta_W)(H_0 + \Delta_H) = W_0H_0 + W_0\Delta_H + \Delta_W H_0 + \Delta_W\Delta_H$$

And

$$\begin{aligned} X - WH &= W_0H_0 + N - (W_0H_0 + W_0\Delta_H + \Delta_W H_0 + \Delta_W\Delta_H) \\ &= N - (W_0\Delta_H + \Delta_W H_0 + \Delta_W\Delta_H) \end{aligned}$$

We can see that

$$\|E\|_F = \|N - (W_0\Delta_H + \Delta_W H_0 + \Delta_W\Delta_H)\|_F$$

That is, if we obtain a good approximation to W_0 and H_0, then Δ_W and Δ_H are small and hence the term $W_0\Delta_H + \Delta_W H_0 + \Delta_W\Delta_H$ are small and we have $\|E\|_F \approx \|N\|_F$.

Measuring the performance of the factorization

Given $X = X_0 + N = W_0 H_0 + N$ and a factorization WH, we now have four different performance measures

$$\|X - WH\|_F \tag{1}$$
$$\|X_0 - WH\|_F \tag{2}$$
$$\|W_0 - W\|_F \tag{3}$$
$$\|H_0 - H\|_H \tag{4}$$

(1) is an approximation to the size of the noise and it will be always greater than zero.
(2) is the amount of difference between X_0 (the "clean" part of X) and the WH (the approximator). It is possible that (2) equals to zero.
(3) and (4) are the distance between the approximated factor matrix and the ground truth. Both (3) and (4) can be equal to zero.
For algorithm comparisons using synthetic data, W_0, H_0 are available, hence (3) and (4) are more preferred.
In practise, W_0, H_0 (and hence X_0) are unavailable, so usually (1) is used.

The non-uniqueness of factorization ... 1

In factorization problems, it is important to know that factorization solution can be "non-unique".

Scalar case example. Given $X = 8 = 2 \times 4$ with $W_0 = 2$, $H_0 = 4$ and there is no noise. Assume we have a factorization algorithm $\mathcal{A}$ that factorizes X into integer W and H. Then all the possible outcome of the factorizations of X are

- $8 = 1 \times 8$, with $W = 1$, $H = 8$
- $8 = 2 \times 4$, with $W = 2$, $H = 4$
- $8 = 4 \times 2$, with $W = 4$, $H = 2$
- $8 = 8 \times 1$, with $W = 8$, $H = 1$

All the 4 factorizations give zero error in terms of $\|X - WH\|_F$, but only the second factorization gives zero error in terms of $\|W_0 - W\|_F$ and $\|H_0 - H\|_F$.

The non-uniqueness of factorization ... 2

In matrix case, such "non-uniqueness" comes from permutations and scaling. Assume we have $X = W_0 H_0$ (with no noise) and an algorithm $\mathcal{A}$ that factorizes the matrix X. The possible factorizations are

- $X = WH$ with $W = W_0$, $H = H_0$
- $X = WH$ with $W = W_0 P$, $H = P^{-1} H_0$ where P is a permutation matrix
- $X = WH$ with $W = W_0 PD$, $H = D^{-1} P^{-1} H_0$ where P is a permutation matrix and D is a diagonal scaling matrix
- $X = WH$ with $W \neq W_0$, $H \neq H_0$ and there is no P or D such that $W = W_0 PD$ and $H = D^{-1} P^{-1} H_0$

Case 1 is perfect factorization as

$$\|X - WH\|_F = \|W_0 - W\|_F = \|H_0 - H\|_F = 0.$$

Case 2 and 3 are in fact case 1 but the columns and rows in W and H are permuted and scaled. In this case direct computation of $W_0 - W$ is meaningless. Case 4 means the solution form the algorithm is a different local minima.

Matching the columns

Consider case 2 : we have $X = W_0 H_0$ with no noise, an algorithm $\mathcal{A}$ that factorizes the matrix X and gives $W = W_0 P$ and $H = P^{-1} H_0$ where P is some unknown permutation matrix.

Since there is a permutation, so the error on W should not be computed directly as $\|W_0 - W\|_F = \|W_0 - W_0 P\|_F = \|W_0(I - P)\|_F$ but instead $\|W_0 - W\hat{P}\|_F$, where $\hat{P} \in \mathbb{R}^{r \times r}$ is the estimated of the inverse of the unknown P.

Fact : permutation matrix are orthogonal. So the inverse of P is $P^\top$.

Hence, if we can estimate $P^\top$ well by $\hat{P}$, we can cancel out the effect of permutation and we have

$$\|W_0 - W\hat{P}\|_F = \|W_0 - W_0 P\hat{P}\|_F = \|W_0 - W_0 \underbrace{PP^\top}_{I}\|_F = \|W_0 - W_0\|_F = 0$$

And the same applies to the rows of H.

The question is : how to compute $\hat{P}$?

Estimation of the permutation matrix by correlation

The easiest way to find $\hat{P}$ is to use correlation.

Consider we have W_0 and W that $W_0 = WP$ with some unknown P. The matrix $\hat{P}$ can be computed as follows :

1. Construct a correlation matrix C that $C_{ij} = \langle W_0(:, i), W(:, j) \rangle$, the correlation between column i of W_0 and column j of W
2. Use C to construct P by looking at the element with largest magnitude.

However, such approach is not robust.

The assignment problem

In fact, $\hat{P}$ can be computed by solving an integer programming problem known as the *assignment problem* : first, construct a matrix C that

$$C_{ij} = \|W_0(:, i) - W(:, j)\|_2$$

i.e., C_{ij} tells the distance between the i^{th} column of W_0 and the j^{th} column of W.
Then, solve the following problem :

$$\begin{aligned} \min \quad & \sum_{i=1}^{r}\sum_{j=1}^{r} C_{ij}X_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{n} X_{ij} = 1 \\ & \sum_{j=1}^{n} X_{ij} = 1 \\ & X_{ij} \in 0, 1 \end{aligned} \tag{5}$$

The assignment problem

Given C, the problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{r}\sum_{j=1}^{r} C_{ij}X_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{n} X_{ij} = 1 \\ & \sum_{j=1}^{n} X_{ij} = 1 \\ & X_{ij} \in 0, 1 \end{aligned}$$

is known as the assignment problem. It is a linear integer programming problem as X_{ij} are restricted to be 0 or 1.

Once we solve such problem, the resulting matrix X will be the permutation matrix $\hat{P}$ we want.

How to solve such problem : the famous Hungarian algorithm.

Last page - summary

Summary :

- Residue / error $E = X - WH$
- Error corresponding to ground truth factors Δ
- For $X = W_0 H_0 + N$, $\|E\|_F = \|N - (W_0 \Delta_H + \Delta_W H_0 + \Delta_W \Delta_H)\|_F$
- Column matching by correlation and the assignment problem

Not included : how to solve the assignment problem.

End of document