An Online Semi-NMF Algorithm for Soft-Clustering of Financial Institutions*

Yuan Cheng
Cornell University
yc956@cornell.edu

Shawn Mankad
Cornell University
spm263@cornell.edu

ABSTRACT

In this paper we develop and propose an online semi-non-negative matrix factorization framework to cluster firms by their stock returns. The model is motivated by an accounting balance sheet identity, where one of the estimated matrix factors can be seen as the percentage of holdings across different asset classes (stocks, bonds, etc.) for each firm – an important input for risk analysis. We also show that our model is an extension of soft K-means clustering. To enhance the practical value of the proposed model (OSNMF), we also develop a fast estimation framework that can be readily applied to cluster firms in real-time as new data becomes available. The model is validated using synthetic and real data. Specifically, we apply our technique to recover asset holdings of mutual funds and ETFs from stock returns and show our estimates closely match their disclosed balance sheets.

CCS CONCEPTS

• Information systems → Clustering.

KEYWORDS

online learning, matrix factorization, financial data

ACM Reference Format:

Yuan Cheng and Shawn Mankad. 2019. An Online Semi-NMF Algorithm for Soft-Clustering of Financial Institutions. In *The 5th Workshop on Data Science for Macro-modeling with Financial and Economic Datasets (DSMM'19)*, June 30, 2019, Amsterdam, Netherlands. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3336499.3338005>

1 INTRODUCTION

In this paper we develop a novel online matrix factorization technique that can be used to cluster firms by the asset-side of their underlying balance sheet. In particular, we build on popular accounting models of financial institutions to motivate a constrained non-negative matrix factorization problem that can be used to infer firm asset holdings (stocks, bonds, etc.) from daily stock returns.

Our factorization problem has two differentiating features when compared to the standard non-negative matrix factorization (NMF;

[8]), where one decomposes a matrix of non-negative entries into lower rank non-negative factors ($X \approx UV^T$). First, only one of the factors in our problem is constrained to be non-negative, that is, the other factor can be composed of elements of any sign. This extension to the NMF factorization, first explored in [5] with the so-called Semi-NMF, allows our model to be potentially useful in other contexts where the input matrix is of mixed signs. The second differentiating feature to our problem is the time-series dimension, which requires that the model be estimated over rolling windows as new stock returns data becomes available each day. Thus, the focus of this work is an *online* semi-NMF, which, to our knowledge, represents a novel development to the online learning and matrix factorization literature.

Our estimation approach builds upon the literature for projected gradient methods for NMF [9] and online NMF [15]. We propose an alternating minimization approach for this problem with projected gradient descent to enforce nonnegativity of one of the matrix factors. The model also has an important relation to the classic unsupervised learning algorithm of K-means clustering [7]. Specifically, we show that the OSNMF is a relaxation of soft or fuzzy K-means clustering [2] that utilizes information about the angle between clusters in addition to distance when clustering observations.

To validate our model and estimation frameworks with real data, we apply the factorization on simulated and real data. The real dataset is composed of the public returns data of the 24 largest mutual funds and exchange traded funds (ETFs). We estimate asset allocations and compare against the firms' true holdings which are obtained through official filings with the US Securities and Exchange Commission and summarized on popular financial websites, like Yahoo Finance.

The paper is organized as follows. Section 2 presents the accounting framework and our rolling window setup. In Section 3, we propose our online semi-NMF method. Section 4 shows the performance of OSNMF on both the synthetic and real datasets. We then conclude with Section 5 summarizing the performance and novelty of our proposed framework and future work.

2 MODEL DEVELOPMENT

Building on previous accounting and factorization models [3, 4], we start by motivating the matrix factorization model through an accounting framework. Let there be n firms under consideration. W_{ik} is the percentage weight invested in each of k assets (equities, bonds, commodities, etc.) by fund i ($\sum_k W_{ik} = 1$), Y_{ikt} denotes the market value of asset k on day t for firm i . E_{it} indicates the market value of fund i 's equity on day t , and D_{it} be the total value of debt liabilities of fund i on day t . A standard balance sheet decomposition would yield $\sum_k W_{ik} Y_{ikt} = E_{it} + D_{it}$, where assets equal liabilities.

*This material is based upon work supported by the National Science Foundation under Grant No. 1633158.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DSMM'19, June 30, 2019, Amsterdam, Netherlands

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6823-0/19/06...\$15.00

<https://doi.org/10.1145/3336499.3338005>

Then define $V_{ikt} = \frac{Y_{ikt+1} - Y_{ikt}}{Y_{ikt}}$, so that $\sum_k W_{ik} V_{ikt}$ is the one day ahead fractional rate of return of the assets held by fund i . By the standard accounting identity,

$$\sum_k W_{ik} V_{ikt} = \frac{(E_{it+1} + D_{it+1}) - (E_{it} + D_{it})}{(E_{it} + D_{it})}. \quad (1)$$

Recall that D_{it} represents debt claims on the firms. D_{it} is often relatively small in size and evolving very slowly. For example, in the context of banks, [6] and [4] assume that debt is time-invariant when interbank debt is separated and observable. For mutual funds this is especially the case, since they are severely constrained by law in their ability to issue debt [11]. In other words, such funds typically issue common stock (equity) to pay for any assets. Thus, we assume that the debt liabilities are small and vary slowly compared to equity. We can then rewrite the expression above as

$$\sum_k W_{ik} V_{ikt} = \frac{E_{it+1} - E_{it}}{E_{it}}, \quad (2)$$

the one day ahead return on equity. Letting $Z_{it} = \frac{E_{it+1} - E_{it}}{E_{it}}$ and assuming that the investment opportunity set is the same for all funds, the underlying factorization problem is

$$\min_{W, V} \|Z - WV^T\|_F^2 \text{ s.t. } W \in \mathbb{R}^{n \times K}, V \in \mathbb{R}^{T \times K},$$

with W subject to probability constraints on its rows.

To gain further intuition about OSNMF, we note that the task of inferring percentage asset holdings for each fund can be viewed as a fuzzy or soft clustering problem. Indeed, one interpretation of our model is that rows of W estimate the posterior probability of belonging to each cluster and columns of V capture the cluster's mean in p dimensional space. In fact, we show that the OSNMF in (3) is a relaxation of fuzzy K-means clustering that utilizes a mixture of Gaussian components with cluster-specific covariances. We start by writing $Z = [z_1, \dots, z_n]^T$ and $V = [v_1, \dots, v_K]$. Then we can rewrite the main objective function as

$$\begin{aligned} \|Z - WV^T\|_F^2 &= \sum_{i=1}^n \|z_i - \sum_{k=1}^K w_{ik} v_k\|_2^2 \\ &= \sum_{i=1}^n \left\| \sum_{k=1}^K w_{ik} (z_i - v_k) \right\|_2^2 \text{ since } \sum_k w_{ik} = 1 \\ &= \sum_{i=1}^n \sum_{k=1}^K w_{ik}^2 \|z_i - v_k\|_2^2 \\ &\quad + \sum_{i=1}^n \sum_{k \neq l} w_{ik} w_{il} (z_i - v_k)^T (z_i - v_l). \end{aligned} \quad (3)$$

Note the first term of (3) is equivalent to the objective function for fuzzy K-means clustering [2, 7] with squared probabilities denoting the strength of association between each observation and cluster. In the second term of (3), if the cluster assignment beliefs are proportional to the distance from the data point to the cluster mean, $w_{ik} \propto \frac{1}{\|z_i - v_k\|}$, then the second term measures the cosine of the angle between clusters when centered on the observed data, i.e., $w_{ik} w_{il} (z_i - v_k)^T (z_i - v_l) \propto \frac{(z_i - v_k)^T (z_i - v_l)}{\|z_i - v_k\| \|z_i - v_l\|} = \cos(\theta)$. OSNMF clusters data by using the correlation between clusters in addition to the distance of observations to cluster means, thus relaxing fuzzy K-means clustering. The implicit reliance on roughly

Method	Mixed Factors	Online
OSNMF	✓	✓
FKmeans	✓	
Online NMF		✓

Table 1: Algorithm Comparison

Figure 1: Rolling Window Setup

Gaussian shaped clusters is appropriate for our context, given that returns are known to be approximately Gaussian with heavier tails [14]. To summarize, Table 1 shows a comparison of our model with related algorithms.

3 ESTIMATION

3.1 Relaxing the Time-Invariant W Assumption

While it was critical to assume that W is fixed over intervals while developing the model to maintain the probability constraints, here we relax the assumption of a fixed W through an estimation trick of using a *rolling window*. As shown in Figure 1, the factorization model in (3) is solved using T days of consecutive data and updated as new observations become available. Specifically, the current percentage weights W_t and asset change V_t of the t^{th} interval are estimated with stock returns Z_t of the current and previous $T - 1$ days. Thus, with this rolling window framework, W can be reported with a daily frequency, while maintaining the probability constraints that are desirable for interpretation.

3.2 Online Estimation Algorithm

To perform the estimation each day in a timely fashion, we adopt an online learning framework. The specific algorithms we develop build on the extant online NMF literature [5, 9, 15]. We introduce terminology and notation to help with exposition. First, let a mini batch represent a subset of data that is used to estimate key intermediate parameters, such as the gradient. Note that by using a subset of data instead of the full dataset, the gradient and thus the overall factorization can be estimated in a computationally efficient matter. We denote the mini batch with $Z^{(t)}$. We also use the term “data input” to refer to one new daily stock returns observation or multiple new days stock returns observations for all funds. Thus, it is a matrix of size n by number of days. We input a mini batch of data for each time into our semi-NMF solver and obtain the estimation $W_S^{(m)}$ after S iteration of update. Then we update $V^{(m+1)}$ with $W^{(m)}$ fixed. Such alternative algorithm online semi-NMF framework is shown in Algorithm 1.

Algorithm 1 Online Semi-NMF (OSNMF) Framework

```

Initialize  $W^{(0)}$  with Fuzzy K means.
for  $m = 1 : M$  do
  Input a set of data  $Z^{(m)}$ .
  Compute the optimal  $V^{(m)}$  with  $W^{(m-1)}$ 
  for  $s = 1 : S$  do
    Compute the gradient  $\Delta_s = \sum_{h=1}^m Z^{(h)}(V^{(h)})^T -$ 
 $W_{s-1}^{(m)} \sum_{h=1}^m V^{(h)}(V^{(h)})^T$ 
    Update the cluster  $W_s^{(m)}$  by
 $W_s^{(m)} = \Delta_s \text{diag}^{-1} \{ \sum_{h=1}^m V^{(h)}(V^{(h)})^T \} + W_{s-1}^{(m)}$ 
    Rotate  $W_s^{(m)}$  to have row sum equal to one by multiplying
    it with a diagonal matrix  $R$ .
  end for
end for

```

We implement an alternating minimization approach. We alternately minimize Equation 4 and 5. The optimal V is estimated by minimizing the following least square problems with $W^{(m-1)}$ fixed.

$$\min_{V^{(m)}} \|Z^{(m)} - W^{(m-1)}(V^{(m)})^T\|_F \quad (4)$$

for m^{th} data input, where Z^m is of size n by r , r is the number of inputting data points. W is the optimal solution with $V^{(m)}$ fixed as a nonnegative least square problem

$$\min_{W^{(m)}} \|Z^{(m)} - W^{(m)}(V^{(m)})^T\|_F, (W^{(m)})_{i,j} \geq 0 \text{ for all } i, j, \quad (5)$$

which we solve with a second-order projected gradient descent

$$x_{s+1} = \Pi[x_s - \Delta f(x_s)H^{-1}\{f(x_s)\}] \quad (6)$$

$H^{-1}(f(x_s))$ is the inverse of the Hessian matrix, which we approximate by Diagonal Approximation instead of exact calculation for computational efficiency. Plugging in W we have

$$W_{s+1}^{(m)} = \Pi[W_s^{(m)} - \Delta_{W_s^{(m)}} f(W_s^{(m)})H^{-1}\{f(W_s^{(m)})\}] \quad (7)$$

where

$$f(W_s^{(m)}) = \sum_{h=1}^m \|Z^{(h)} - W_s^{(m)}V^{(h)}\|_F^2 \quad (8)$$

We choose the second-order projected gradient descent (PGD) instead of the first-order PGD for a faster convergence as the latter could be costly when searching for the right step size [9]. Finally, as results of matrix factorization method usually vary a lot with different seeding techniques. To improve the stability, we seed the algorithm with cluster centers of Fuzzy K means.

4 VALIDATION AND RESULTS

4.1 Synthetic Data Generation

Here we generate data according to the model implied by OSNMF to show that the estimation method is self-consistent. We simulate W and V according to distributional assumptions, and subsequently obtain corresponding $Z = WV$. We assume W follows a Dirichlet distribution (with hyperparameter $\alpha = 0.5$), which automatically satisfies the probability constraints on the rows of W . For V , we each element is generated by a normal distribution with zero mean and variance equal to 0.01. We hold W fixed and let V vary to mimic

the real-data setting, where the funds/ETFs in our data are unlikely rebalance their portfolios often or engage in active trading. The other parameters (input matrix dimension, desired number of asset classes, number of periods, and period length) for generating the simulated data are set to match the real dataset, which is described next.

4.2 Real Data Description

Recall our goal here is to use the daily closing prices of mutual funds/ETFs (hereafter referred to as funds) to infer their actual percentage asset holdings. Let $Z_{it} = \log(\frac{p_{it}}{p_{it-1}})$, where p_{it} is the closing price of fund i on day t , that is, we construct $Z = [Z_1, \dots, Z_T]$ using daily log returns. The total duration of our data is 49 trading days, ending with the day prior to official fund disclosures to the Securities and Exchange Commission detailing their major asset holdings. We have $n = 24$ funds with 25 rolling windows if the interval length $p = 25$ days. This translates to 25 daily estimates of W and V with number of asset classes $K = 3$.

Though we estimate both W and V , we focus on the accuracy of W to assess clustering accuracy. Note that we have the true percentage asset holdings W for these funds from their official disclosures to the SEC as of June 30, 2017. Their percentage asset holdings in US Stocks, non-US stocks, and bonds. Note that a handful of funds report holding cash or "other" assets. Arguably these could be additional asset classes to consider but because they typically represent less than 1% of overall holdings, we ignore them here.

4.3 Results

We assess the performance of each method in several ways. First, we assess the accuracy of the estimated W by comparing it to the true W with element-wise RMSE. The average Frobenius loss of the Z and W in Table 2 shows the advantage of OSNMF over Fuzzy K means. OSNMF achieves a more accurate estimate of the true W averaging over all the rolling window periods. In general, OSNMF outperform Fuzzy K means in terms of individual rolling window. As W can be viewed as a clustering solution, we also calculate the adjusted Rand Index (ARI; [12]) using the nearest hard clustering of both the estimated and true W . When ARI equals one, it means the two clusterings are identical. From Table 2, the ARI is high for both the simulated and real datasets. This shows the ability of OSNMF in recovering clustering information.

Driven by our application, where the distribution of elements in W have risk implications, we also compare the true and estimated W with two distributional tests. The first distributional test is the Mann-Whitney U test [10] to assess whether our estimate of W is stochastically smaller (or larger) than its true value, i.e., let x and y be two random variables with cumulative distribution functions f and g respectively. The hypotheses for the test is then $f = g$ versus the alternative hypothesis that $f(a) < g(a)$ or $f(a) > g(a)$ for all a . From the results in Table 2, our hypothesis for the Mann-Whitney U test is not rejected at a 5% significance level, i.e., the hypothesis holds that a randomly selected sample in the distribution of estimated W equals a randomly selected sample in the distribution of true W . The second and more stringent test we utilize is the Two Sample Anderson Darling Test, created by [13] based on the classical Anderson Darling Test [1]. The idea here is to assess

Dataset	$\ Z - Z_{OSNMF}\ _F$	$\ W - W_{OSNMF}\ _F$	$\ W - W_{FKMeans}\ _F$	ARI_{OSNMF}	MWU_{OSNMF}	AD_{OSNMF}
Simulation	0.0012 (0.0001)	0.1443 (0.0068)	0.1729 (0.0059)	0.7229 (0.0934)	0.0612 (0.0175)	0.0004 (0.0003)
Mutual fund & ETF	0.0013 (0.0001)	0.0949 (0.0265)	0.1022 (0.0139)	0.9129 (0.0880)	0.5100 (0.2355)	0.0428 (0.0420)

Table 2: Performance Results of the proposed Online Semi-NMF (OSNMF) in comparison with Fuzzy K means (FKMeans), the numbers in paranthesis are standard deviation from 1000 instances. For Mann-Whitney U test (MWU) and Anderson Darling Test (AD), corresponding p-value and standard deviation are reported.

Figure 2: Distribution of W from OSNMF and Distribution of True W for Mutual Fund/ETFs

whether there are differences between the two samples with particular sensitivity at the tails of the sampled distributions. With both tests, failing to reject the null hypothesis would provide statistical evidence of the validity of the model and estimation procedure. The results from Table 2 indicate rejection of this hypothesis, i.e., the tails of the two sample distribution are different. The empirical cumulative density functions in Figure 2 also give evidence of this difference in the sample tails.

5 CONCLUSIONS

We have developed and proposed an online semi-non-negative matrix factorization framework for estimating the percentage asset holdings of firms from their stock returns. Our projected gradient descent and rolling window estimation framework provides for fast, sequential (daily) updating. We validated our proposed algorithm on both synthetic and real data. In all of our experiments, the proposed algorithm outperformed Fuzzy K means, a standard method for soft clustering, in terms of both element-wise RMSE and the adjusted rand index. Thus, the proposed algorithm works well in terms of hard and soft clustering accuracy, although the tail distribution of the estimated factor weights differs from the true distribution on the real data. A natural next step for future work is to apply the estimation framework on data from other financial institutions, such as banks, and to construct risk measures from the estimated W .

REFERENCES

- [1] Theodore W Anderson and Donald A Darling. 1954. A test of goodness of fit. *Journal of the American statistical association* 49, 268 (1954), 765–769.
- [2] James C Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10, 2-3 (1984), 191–203.
- [3] Celso Brunetti, Jeffrey H Harris, and Shawn Mankad. 2018. Bank Holdings and Systemic Risk. *Finance and Economics Discussion Series* 2018, 063 (Aug. 2018), 1–52.
- [4] Celso Brunetti, Jeffrey H Harris, Shawn Mankad, and George Michailidis. 2019. Interconnectedness in the interbank market. *Journal of Financial Economics* (Jan. 2019).
- [5] C. Ding, Tao Li, and M.I. Jordan. 2010. Convex and Semi-Nonnegative Matrix Factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32, 1 (2010), 45–55. <https://doi.org/10.1109/TPAMI.2008.277>
- [6] Matthew Elliott, Benjamin Golub, and Matthew O Jackson. 2014. Financial networks and contagion. *American Economic Review* 104, 10 (Jan. 2014), 3115–3153.
- [7] Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.
- [8] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (10 1999), 788–791.
- [9] Chih-Jen Lin. 2007. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural computation* 19, 10 (Oct. 2007), 2756–2779.
- [10] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [11] John Morley. 2013. The Regulation of Mutual Fund Debt. *Yale J. on Reg.* 30 (2013), 343.
- [12] William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66, 336 (1971), 846–850.
- [13] Fritz W Scholz and Michael A Stephens. 1987. K-sample Anderson-Darling tests. *J. Amer. Statist. Assoc.* 82, 399 (1987), 918–924.
- [14] DAVID E UPTON and DONALD S SHANNON. 2012. The Stable Paretian Distribution, Subordinated Stochastic Processes, and Asymptotic Lognormality: An Empirical Investigation. *The Journal of Finance* 34, 4 (April 2012), 1031–1039.
- [15] Fei Wang, Chenhao Tan, Arnd Christian König, and Ping Li. 2011. Efficient Document Clustering via Online Nonnegative Matrix Factorizations. *Proceedings of the SIAM International Conference on Data Mining* (2011), 908–919.